缶蹴りにおける他者知覚を必要としない協調の実現

Cooperation without Person Perception on Kick the Can Game

高田 亮介 [1*]　坂本 孝丈 [1]　竹内 勇剛 [1]
Ryosuke Takata[1]　Takafumi Sakamoto[1]　Yugo Takeuchi[1]

[1] 静岡大学
[1] Shizuoka University

Abstract: 人と機械の関係は “人が機械を使う” という道具的な関係から “人と機械が協調する” という社会的な関係に変化してきている．人が他者と協調するためには，同一の目標に向けて相互に行動を調整する必要がある．本研究では缶蹴りを題材とし，ボトムアップに協調的な振る舞いを構築する強化学習を用いて，敵・味方の両方が相互適応して振る舞いを変化させるシミュレーション実験を行った．実験の結果，他者知覚を必要としなくても，1 対 1 のインタラクションに対する報酬系がエージェント間の協調的な関係を創発させることが示唆された．

1　はじめに

コロナ禍によって遠隔コミュニケーションの機会が増加し，他者の行動を知覚できない状況で協調的に振る舞うことが求められるようになった．このような日常的に行われる他者との協調について理解するうえで，協調的な振る舞いが可能なエージェントを実現し分析する構成論的手法が有用である．本研究で扱う協調とは “同一の目標に向けて行動を調整するインタラクション” のことを示し，協調系は複数のエージェントによる創発現象と捉えることができる．マルチエージェント系を用いた構成論的手法によって，協調的な振る舞いが創発し得る条件，およびその過程を明らかにできれば，人と協調可能な人工物の設計や人同士の協調を促すコミュニケーションシステムの構築に寄与し得る．

これまで，エージェントを用いて協調的なインタラクションを実現する研究は多く行われてきた．マルチエージェント系では，個々のエージェントが協調的に振る舞うことによって全体の作業効率を向上させる分散人工知能（Distributed Artificial Intelligence, DAI）の研究 [1] や，生物の群れの挙動をモデル化し，衝突を避けつつ同じ方向に進む振る舞いのシミュレーション [2] が行われてきた．また，人は他者に対して協調的に振る舞うことが可能であり，人とエージェントの間においても協調的なインタラクションが成立し得る．例えば，竹内らは人の持つ社会性によってエージェントとの協調的なインタラクションが実現可能であることを示唆し [3]，中嶋らはあらかじめ決められた社会的応答を実行するエージェントに協調的なインタラクションが見出せることを示した [4]．以上の研究は，トップダウンに社会的なインタラクションをモデル化し，協調的な振る舞いを見出している．このトップダウンな方法は，個々のエージェントの行動戦略およびエージェント間のコミュニケーション規約を事前に知識としてシステムに組み込むことを前提にしている．これらの知識は問題領域に依存することから，知識獲得の問題に直面する [5]．この問題を解決するためには，エージェントのモデル化をボトムアップに行う必要がある．

ボトムアップに協調系を創発させる手法として，しばしば強化学習が用いられる [6]．椿本らはグリッド空間課題，Raileanu らは単純な選択課題に対して，意図推定をモデル化した強化学習によって円滑に協調できることを示した [7, 8]．大倉らは，連続空間の協調課題に対してマルチエージェント強化学習を適用し，群知能的な協調が創発することを示した [9]．これらの強化学習を用いて協調課題を解く先行研究は，学習の途中段階の分析を行っていないため，どの学習段階でどのような振る舞いを獲得したかについて言及されない．個々のエージェントの創発現象である協調系について理解するためには，どの学習段階でどのような振る舞いが獲得されたか，といった学習過程の分析を行う必要がある．

近年は協調的な振る舞いを実現するために必要な要素について，他者知覚の有無に注目した研究が行われている．Uwano らは，他者を含む環境情報を観測しない強化学習手法を用いて，グリッド空間上の単純な課題で協調的な振る舞いの実現に成功した [10, 11]．他者知覚を行わないことは，他者モデルを明示的に持たないということであり，協調系を創り出す際に他者モデルが必要ない状況があるということを示唆している．こ

*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: takata.ryosuke.18@shizuoka.ac.jp

のように，協調課題の構造次第では他者知覚を行うことなく協調が実現可能であると示すことは，協調系を創発し得る環境の多様性を向上させることに繋がる点で重要である．他者知覚を行うことなく協調的な振る舞いが求められる課題において，協調対象のエージェント群と敵対する存在が環境上にいて，共適応しながら協調的な振る舞いを獲得していく状況は，生態学上および実環境上では頻繁に起こり得る．このような状況を満たす課題での研究は，Reynolds による鬼ごっこ [12] や，Tampuu らによる Pong ゲーム [13]，Baker らによるかくれんぼ [14] といった課題で行われているが，敵対するエージェントを参照することで味方を直接知覚しなくてもよい，という構造への言及はまだ行われていない．

本研究では，マルチエージェント強化学習による協調的な振る舞いの学習において，他者の知覚の可否が及ぼす影響を学習過程の分析を通して検証する．題材として，敵が存在する競争課題でありながら，味方同士が協調することで課題を有利に進められる協調課題でもある缶蹴りを用いる．缶蹴りは，集団レベルでの協調系をミクロに見るとエージェントの 1 対 1 のインタラクションになっているという構造で，共適応によって適応的戦略が一意に定まらないという特徴を持つ．実験条件として，他者を知覚可能な状況と知覚不可能な状況，さらに利己的な振る舞いを誘発する報酬系を用いた分析を行う．これにより，他者知覚を行うことなくエージェント間で協調するために振る舞いの調整が行われること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．本研究の成果は，他者と直接的なコミュニケーションすることなく協調可能かどうかを議論することに繋がり，他者を直接知覚しない遠隔コミュニケーションの構造の解明に寄与し得る．

2　缶蹴り

2.1　缶蹴りの特徴

缶蹴りは，鬼と子という敵同士が同じ環境に存在する競争的な課題でありながら，子同士の協調によってゲームを有利に進めることができる協調的な課題である．さらに，主な観察対象である子だけでなく，鬼も強化学習によって子に適応することで，鬼と子の間に適応関係が生まれる．すなわち，図 1 のように子が鬼の戦略に対して適応的な戦略を創発し，鬼は子が創発した適応的な戦略に対してさらに適応する，という構造になっている．このとき，子は鬼との 1 対 1 のインタラクションを行い，鬼からのフィードバックにより，自身の振る舞いを修正する．子と鬼との 1 対 1 の勝負では子が勝つことは不可能だが，子と鬼とのインタラクションが別の子と鬼とのインタラクションに影響を及ぼすことで，子が勝利することが可能となる．すなわち，鬼から与えられるフィードバックを介して子の間に協調系が創発すると考えられる．本研究では，このような状況下で子が他者知覚を行わなくても協調可能であることを検証する．

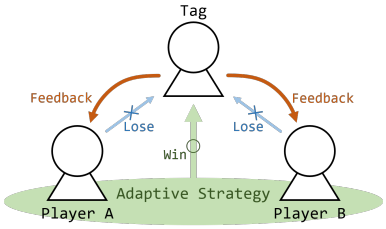


図 1: 缶蹴りの構造（子と鬼とのインタラクションが他の子と鬼とのインタラクションに影響することで協調系が創発する）

2.2　缶蹴りのルール

缶蹴りの手順は以下の通りである．なお，本研究で注目しない手順については省略している．

(1) 参加者を鬼と子の役に分ける．

(2) 缶をフィールドの中央に配置し，子は隠れる．

(3) 鬼は子を探し，見つけた場合はシグナルを出した後で缶に触れることで，見つかった子は退場する．

(4) 鬼が子を全員退場させたら鬼の勝利でゲームを終了する．子のうち 1 人でも缶に触れたら子の勝利でゲームを終了する．

2.3　缶蹴りにおける協調

ここでは，缶蹴りにおける子の協調的な振る舞いを定義する．子側が勝利するためには，子が 1 体でもゲームの最後まで鬼に発見されずに生き残るか，鬼に捕まることなく缶を蹴る必要がある．鬼に発見されずに生き残る勝利は，鬼が探索行動を採るかどうかに依存するため，子側が能動的に制御することは難しい．すなわち，子にとってゲームを有利に進めるための戦略は，鬼に捕まることなく缶を蹴るための戦略であると言える．今回は，他の子が鬼に発見されずに缶に近づくことを可能にする振る舞いを協調的な振る舞いとする．ある子が鬼に発見されずに缶を蹴るためには，他の子が同時に缶に近づき，自ら鬼に発見される自己犠牲の振る舞いが考えられる．すなわち，“複数の子が同時に缶に近づく振る舞い” が缶蹴りにおける協調的な振る舞いである．

2.4　シミュレーション環境

缶蹴りのシミュレーション環境は Unity で作成した．Unity は 3 次元仮想環境の物理演算が可能で，缶蹴りのようなエージェントの相互作用がもたらす複雑系のシミュレーションに適している [15]．Unity で作成した缶蹴り環境を図 2 に示す．図 2 において，紫色のエージェントが鬼，青色のエージェントが子，白い円柱は壁である．鬼は 1 体，子は 4 体配置した．子の数に関しては，社会集団の中での知性を明らかにするためには 3 体以上のエージェントによる協調が必要である [16] という点と，最も少数である 2 体での協調系が別個体の組み合わせで同時に複数出現可能であるため協調系が創発しやすいと考えられる点から 4 体とした．鬼の初期位置はフィールド中央に配置された缶の直前，子の初期位置はそれぞれフィールド左上，右上，左下，右下とした．なお，子の初期位置は鬼の初期位置からは円柱に阻まれて発見不可能な位置に設定した．フィールドの外周は壁に囲まれており，エージェントが隠れるためのオブジェクトとして円柱を 10 個配置した．

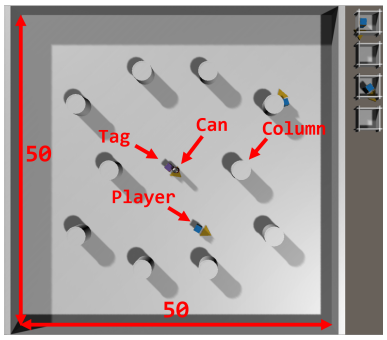


図 2: 缶蹴りシミュレーション環境

シミュレーションに用いる鬼，子は自律移動型のエージェントであり，前方 120 度の視界を有している．図 3 のように，視界内には 11 本の光線を飛ばし，エージェントは光線に当たっているオブジェクトの情報を取得する．

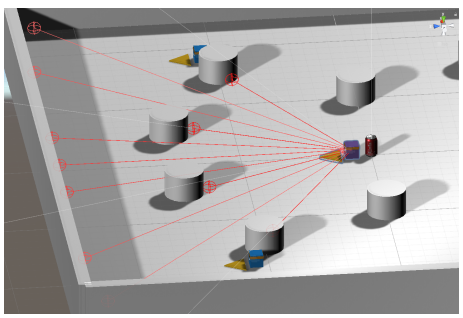


図 3: エージェントの視界と光線

3　強化学習手法

3.1　PPO

強化学習アルゴリズムには Unity ML-Agents に搭載されている PPO (Proximal Policy Optimization)[17] を用いた．PPO 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返す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等で成果を出している [17, 18]．図 4 に，PPO のフローチャートを示す．

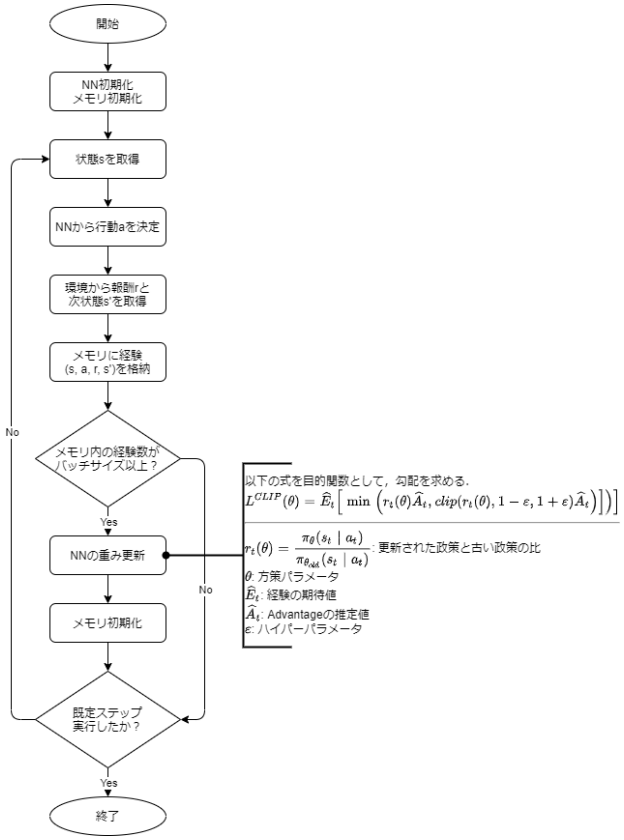


図 4: PPO のフローチャート

PPO の特徴は，方策関数を更新する際に，その変化量が大きくなり過ぎないようにクリッピング操作を行うことで，学習を安定化させている点である．方策の更新は式 (1) を目的関数とした勾配法を用いる．クリッピング操作は式 (1) 中の clip 関数であり，式 (2) に示す方策の変化量比の値が $1-\epsilon$ 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．

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

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

3.2 状態空間とハイパーパラメータ

強化学習におけるエージェントの状態空間は，図 3 のようにエージェントの視界内に飛ばされた 11 本の光線に当たっているオブジェクトの情報と，エージェントの絶対位置情報，発見情報から構成される．発見情報は 2 値をとる変数で，鬼の場合は「子のうち誰かを発見しているか」，子の場合は「自分が鬼に発見されているか」を表す．状態空間を表 1 にまとめる．なお，視界の光線は子を個体識別することが可能であるため，鬼が子を見つけてシグナルを出す（名前を呼ぶ）という缶蹴りのルールを実装できる．

表 1: 状態空間

状態	次元数
エージェントの視界内光線に当たっているオブジェクト情報	11
エージェントの絶対位置座標	3
発見情報（鬼: 誰かを発見したか，子: 自分が発見されたか）	1

PPO におけるハイパーパラメータは表 2 にように設定した．なお，今回の設定は Unity ML-Agents のデフォルト設定を用いた．また，エピソードの最大ステップ数は 2000 ステップとした．

表 2: PPO のハイパーパラメータ

パラメータ名	値
バッチサイズ	128
バッファサイズ	2048
バッファに追加するステップ数	64
方策変化量の閾値 ϵ	0.2
エントロピー正規化率 β	0.005
正規化パラメータ λ	0.95
学習率 η	0.0003
割引率 γ	0.99
エポック数	3
隠れ層のニューロン数	256
隠れ層の数	2
RNN メモリサイズ	128
RNN 経験シーケンス長	64

3.3 報酬によるルールの教示

強化学習で使用する報酬に関しては，2.1 節で述べた缶蹴りのルールにおける勝利条件，敗北条件に従って設定した．鬼の勝利条件かつ子の敗北条件は，鬼が子を発見し缶に触れることであるため，その一連の行動に対して報酬を設定した．子の勝利条件かつ鬼の敗北条件は，子が缶に触れることであるため，その行動に対して報酬を設定した．以上の報酬系を表 3，および表 4 にまとめる．

表 3: 鬼の報酬

内容	値
鬼が（子を発見後）缶に触れる	+1/PNUM
子が缶に触れる	-1
時間経過	-0.0005

表 4: 子の報酬

内容	値
子が缶に触れる	+1
鬼が（子を発見後）缶に触れる	-1
時間経過	+0.0005

学習の初期状態では，学習器であるニューラルネットワークの重みがランダムに設定されるため，最初はエージェントがランダムウォークするだけの状況だが，学習を重ねることで表 3 や表 4 の報酬が最大になるような振る舞いを獲得する．このとき，設計者が与える報酬は缶蹴りのルールを教示しているに過ぎず，エージェント個体が行う協調的な振る舞いは，学習によってボトムアップに創発すると考えられる．

4 学習実験

4.1 実験条件

本研究では缶蹴りを題材に，強化学習によって，他者知覚可能な状況，および他者知覚不可能な状況の両方において協調的な振る舞いを獲得することを検証した．また，利己的な振る舞いを学習し得る条件として，子の間に共通の報酬を与えず個別に報酬を与える状況で実験を行い，協調的な振る舞いの創発において，子の間で報酬を共通にすることの必要性を検証した．そのために，表 5 に示す 4 条件を用意した．ORL 条件は，これまでマルチエージェント強化学習で行われてきた実験と同様に，他者知覚可能な状況で学習させる条件である．本研究では，他者知覚できない状況で協調可能かどうか検証することが目的であるため，他者知覚情報を状態空間に含まない RL 条件の結果が ORL 条件と同程度に協調できているか確認した．さらに，子の報酬を共通にせず，個別に報酬を与える OL 条件および L 条件が子の利己的な振る舞いを誘発するかどうか検証した．

表 5: 実験条件

条件名	共通報酬	他者知覚
ORL	○	○
RL	○	×
OL	×	○
L	×	×

4.2 分析手法

本稿では，強化学習の過程を分析することで，エージェントがどのような行動を獲得していったのか分析する．そのために，エージェントと缶との位置関係から求められるゲームのダイナミクスの変化，およびゲーム終了時のエージェントの状態変化を観察する．

さらに，2 体の子に対する相互相関関数（CCF）を求めることで協調的な振る舞いを評価する．CCF を用いることで，２体のエージェント間の行動の時間的な遅れを検出することができる．子 P_A，子 P_B の缶に対する接近量の時系列データ D_A，D_B に対して，以下の式で表される CCF を適用する．

$$f(k) = \frac{Cov\left(D_A^{(t)}, D_B^{(t+k)}\right)}{\sqrt{(\sigma_{D_A}\sigma_{D_B})}} \tag{3}$$

ただし，t と $t+k$ は時間幅を表し，Cov は共分散，σ_{D_A} は D_A の分散，σ_{D_B} は D_B の分散を表す．これにより，k の値に応じて D_A に対して D_B の時間間隔をずらしたうえでの相関関係を求めることで，D_A と D_B の間の時間遅れを検証することができる．具体的には，D_A と D_B の間にほとんど時間遅れがなく，同時に値が大きくなる場合，$k = 0$ 付近が CCF の最大値になる．D_A の増加に対して D_B があるステップ数 n だけ遅れて増加する場合，$k = n$ 付近が CCF の最大値になる．逆に，D_B が D_A よりもステップ数 n だけ先行する場合は $k < -n$ 付近が CCF の最大値になる．本研究ではエピソード内の缶への接近量を計算し，各子の組み合わせに対して CCF を求める．なお，時間遅れの幅は $-200 \leq k \leq 200$ とする．これにより，一方の子が缶へ接近したタイミングに合わせて，他の子が缶へ接近するような協調的な行動が生じたか否かを検証する．そこから，学習過程に伴うインタラクションの変化を捉える．

4.3 結果と考察

4.3.1 獲得報酬の推移

学習の結果得られた報酬の推移を図 5 に示す．図 5 において，横軸は学習ステップ数，縦軸は獲得報酬を表す．

子の間に共通の報酬を与えた ORL 条件および RL 条件では，鬼と子の獲得報酬が何度も交差している．これは，子だけではなく鬼も学習によって適応するため，鬼と子が共適応の関係になっていることが要因であると考えられる．ORL 条件と RL 条件を比較すると，ORL 条件の方が交差する間隔が広いことがわかる．これは，他者知覚を行う場合は他者知覚を行わない場合よりも，鬼の適応に対して頑健な戦略を学習していることによると考えられる．

子の報酬を個別に与えた OL 条件および L 条件では，鬼よりも報酬を獲得する子としない子に分かれている．このとき，両条件とも学習の初期段階では子の報酬の変化量は同様だが，学習が進むにつれて子の間に獲得報酬の差が生まれていることがわかる．OL 条件および L 条件において，子は自らの報酬を高めるために学習を行うため，利己的な行動を学習したと考えられる．

4.3.2 エピソード長の推移

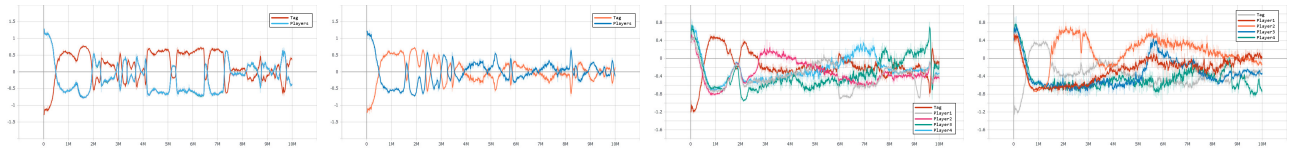
学習過程におけるエピソード長の推移を図 6 に示す．図 6 において，横軸は学習ステップ，縦軸はエピソード長を表し，上から ORL 条件，RL 条件，OL 条件，L 条件のエピソード長の推移を示している．

図 6 より，全ての条件でおよそ 1M ステップまでの間にエピソード長が短くなっていることがわかる．これは，鬼が子を捕まえる振る舞いを学習した，あるいは子が缶を蹴る振る舞いを効率化したことが要因である．また，学習過程全体を通して ORL 条件は他の条件に比べてエピソード長が短い．RL 条件，OL 条件，L 条件では，学習の途中からエピソードが長期化することが多くなっている．子が 1 体でも捕まっていなければエピソードは終了しないため，エピソードが長期化する要因としては，子が鬼に見つからない位置（円柱の後ろ等）に隠れる振る舞いを学習したことが考えられる．

4.3.3 缶への接近量の変化

学習過程の子における缶への接近量の変化を図 7 に示す．図 7 において，横軸は学習ステップ，縦軸は缶への接近量を表し，上から子 P1（左上），子 P2（右上），子 P3（左下），子 P4（右下）の接近量の変化を示している．

図 7 より，缶に近づく振る舞いを学習した子と，缶から離れる，または初期位置から振る舞いを学習した子に分かれている．缶への近づきやすさの条件である初期位置から缶までの距離は，学習全体を通してどの子も同一である．条件が同一であるにも関わらず，学習の中で缶に近づく子と缶に近づかない子に分化したことは，子がダイナミクスを学習していく中で，各々が自身の立ち位置を方向づけていったと考えられる．



(a) ORL 条件　(b) RL 条件　(c) OL 条件　(d) L 条件

図 5: 学習の結果得られた報酬の推移

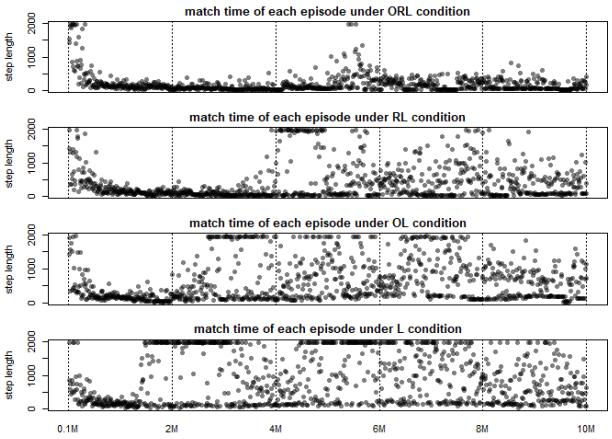


図 6: 学習過程のエピソード長の推移

4.3.4　最終状態の変化

学習過程のエージェントの最終状態の変化を図 8 に示す．最終状態は，鬼は “時間切れ”，“敗北”，“勝利” の 3 状態，子は “生存”，“キック”，“死亡” の 3 状態であり，スライディングウィンドウを 10 にしたときの各ゲーム終了時の状態をカウントした．図 8 において，横軸は学習ステップ数，縦軸は状態カウント値，色は状態の分類を表し，上から鬼，子 P1，子 P2，子 P3，子 P4 の最終状態の変化を示している．

図 8 より，ORL 条件では子 P1 は 1 度も缶を蹴っていないため，缶を蹴る動作を学習していないと考えられる．また，子 P1 以外の 3 体はまとまった期間で缶を蹴る振る舞いが循環していることがわかる．ここで，図 7 より，缶を蹴っていない子も缶に接近していることがわかる．すなわち，子の最終状態は鬼がどの子をターゲットにするかによって変化することがあり，振る舞いの結果缶を蹴ることができたかどうかは学習に影響を与える．つまり，鬼と子の 1 対 1 のインタラクションが学習に大きな影響を及ぼすことが考えられる．

4.3.5　協調的な振る舞いの評価

CCF を用いて学習過程の子の間の同期性の変化を図 9 に示す．図 9 において，横軸は学習ステップ数，縦軸は時間遅れ，色は缶への接近量の相関を表し，上から子 P1–子 P2，子 P1–子 P3，子 P1–子 P4，子 P2–子 P3，子 P2–子 P4，子 P3–子 P4 の CCF の推移を示している．時間遅れが 0 に近く，色が赤に近いほど，2 体の子の振る舞いは同期しており，協調的な振る舞いであると言える．

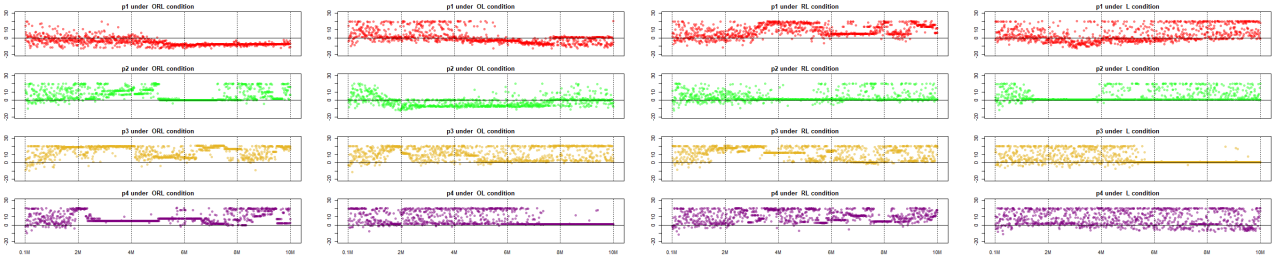
図 9 より，ORL 条件および RL 条件では学習全体を通して微小な時間遅れで相関が強い子が存在することがわかる．それに対して，OL 条件および L 条件では，相関が強いことはあっても時間遅れが大きく，ORL 条件，RL 条件ほど同期性が高いとは言えない．図 7 より，OL 条件と L 条件においても子は缶に接近しているため，同期的ではなく，タイミングをずらして缶に接近していることがわかる．ORL 条件と RL 条件では誰かが缶を蹴ることができれば子全体の利得になるため，同期的に缶に接近することを学習したと考えられる．

5　議論

5.1　鬼と子による軍拡競争

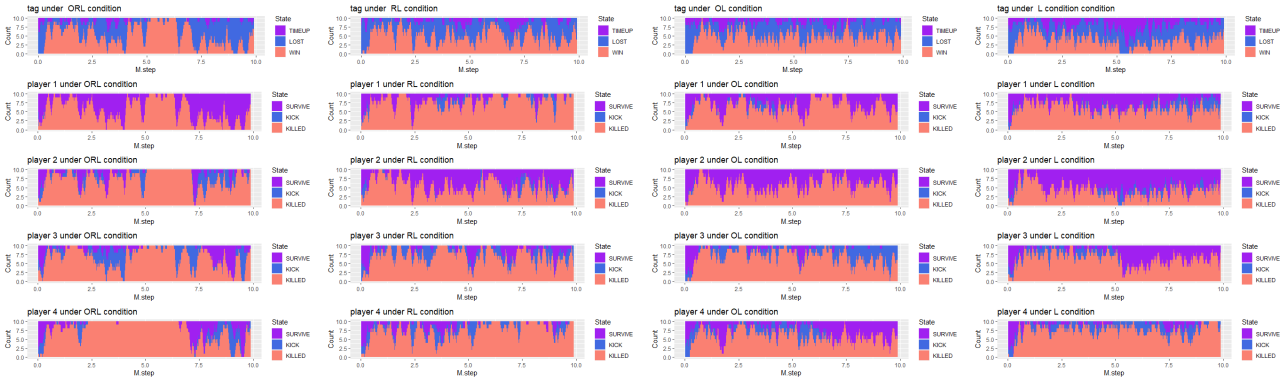
図 5 に示した ORL 条件や RL 条件の学習結果では，鬼と子の報酬が交差しているステップが複数箇所に見られ，10M ステップまでの学習では収束することは無かった．この結果から，図 10 のように，鬼と子は戦略が相互に対応するように振る舞いを循環させながら学習を行っていると考えられる．

このような循環は，Dawkins らが提唱した進化的軍拡競争 [19] に類似した結果である．軍拡競争を収束させるためには，攻撃側と防衛側のリスクを非対称にして，適応的な戦略が一意に定まる課題にする必要がある．すなわち，鬼と子の報酬を非対称にすることで学習の振動が抑えられ，エージェントの戦略に妥協が生まれると考えられる．しかしながら，妥協の戦略を許さず，適応的な戦略が一意に定まらない缶蹴り課題を強化学習させたエージェントが創発する協調にこそ，人と協調可能な人工物の設計や人同士の協調を促すコミュニケーションシステムへの応用可能性を見出せると考える．



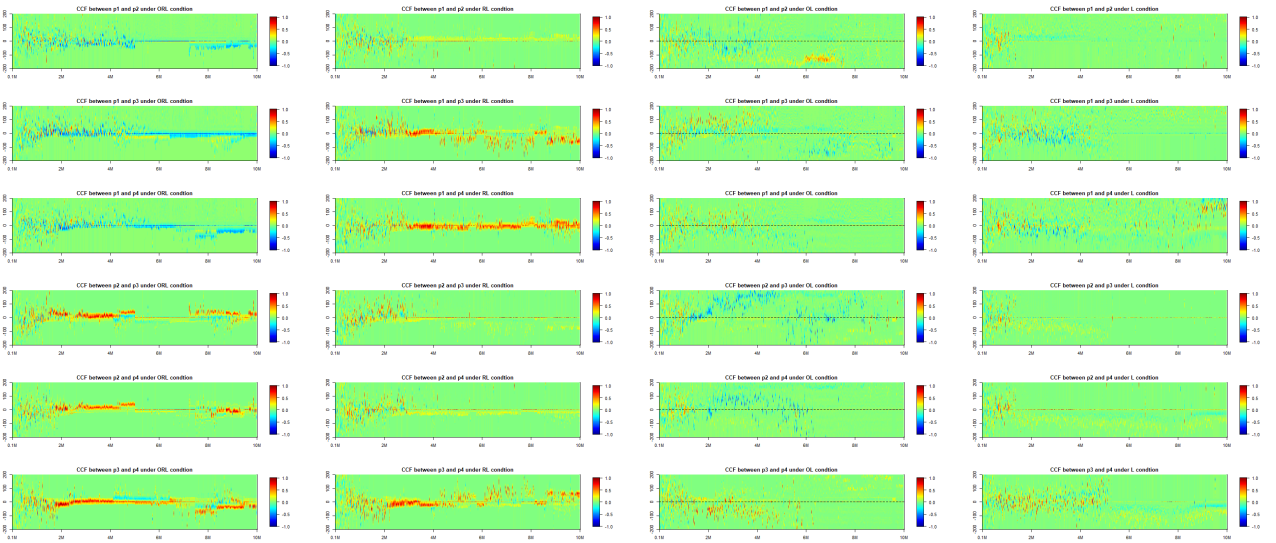
(a) ORL 条件　(b) RL 条件　(c) OL 条件　(d) L 条件

図 7: 学習による子の缶への接近量の変化



(a) ORL 条件　(b) RL 条件　(c) OL 条件　(d) L 条件

図 8: 学習による鬼と子の最終状態の変化



(a) ORL 条件　(b) RL 条件　(c) OL 条件　(d) L 条件

図 9: 学習による子の同期性の変化

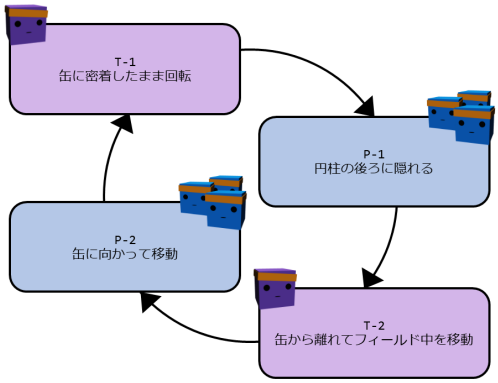


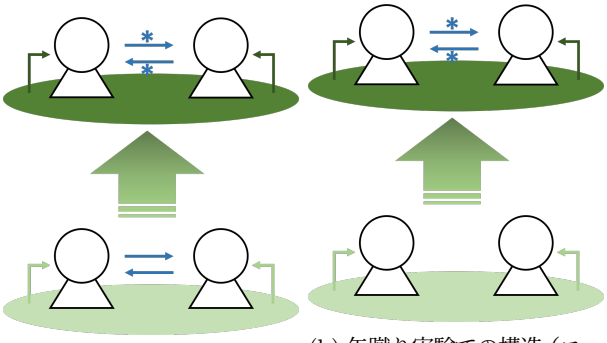
図 10: 振る舞いパターンの循環

5.2 エージェントの視点から見た振る舞い

ここまで，環境を俯瞰的に見た第三者的な視点から考察したが，実際の学習はエージェントの主観的な視点から得られる情報を基に行われる．エージェントが得られる環境の情報は，ORL 条件および OL 条件では自身の絶対位置と自身の視界に入るオブジェクト情報，RL 条件および L 条件では自身の絶対位置のみである．他者が知覚できない条件に限らず，ORL 条件においても，円柱の後ろに隠れた子は実際には知覚されていない状態と同じである．そのため，例えば鬼が左上に隠れた子を見つける振る舞いは，鬼が学習の過程で網羅的にフィールドを探索した結果 “左上に子が隠れている” という経験を積んだことにより獲得したと考えられる．また，子がタイミングを合わせて缶に向かって移動する協調的な振る舞いにおいても，子が学習の過程で網羅的に缶までのアプローチのタイミングを探索した結果，どのタイミングで移動を開始すると報酬が得られやすいか経験し，同時に缶に向かって移動する戦略を獲得したと考えられる．

5.3 他者知覚と協調

本実験によって，報酬を子の間で統一した ORL 条件と RL 条件では協調的な振る舞いが見られ，報酬をこの間で統一せず個別に与えた OL 条件と L 条件では協調的な振る舞いは見られなかった．学習で子が行うことは表 4 に示した報酬を最大化することであり，子は他者を直接観察することなく，報酬によって統一された目標に向かって行動を調整することで協調的な振る舞いが可能となっている．以上のように，ボトムアップなアプローチによって目的を達成するために社会性を形成し得ることが示唆された．協調的な振る舞いを創発させる構造は，従来のような図 11(a) の構造だけで無く，図 11(b) のような直接的なインタラクションの無い構造でも可能であることが明らかになった．



(a) 従来の構造（エージェントは他者の身体位置などを参照しながら行動を予測して調整する）

(b) 缶蹴り実験での構造（エージェントは他者の身体位置などを参照することなく協調の指標（報酬）を基に行動を調整して協調する）

図 11: 協調系がボトムアップに構築される構造

5.4 学習手法の検討

本研究では，PPO を用いてボトムアップにエージェントのモデルを学習させた．PPO を用いることにより，鬼と子による共適応的な相互作用の中で急激に変化するダイナミクスに対して，緩やかに方策を更新することで学習に成功したと考えられる．ここで，PPO のように方策勾配に制限を設けない手法や，ハイパーパラメータの学習率等を大きくした場合を考えると，学習が収束しない可能性がある．急激に方策勾配降下される強化学習手法を用いる，またはハイパーパラメータの設定を変更するなど，学習アルゴリズムに関する項目を変更した場合，適応的な戦略が一意に定まる課題であればその戦略獲得を目指す学習エージェント群にのみ影響が及ぶが，共適応する課題では学習エージェント群が複数存在するためその全てに影響し，さらに影響を受けて振る舞いが変化した学習エージェント群によって他の学習エージェント群の振る舞いが変化する．

5.5 個体数の検討

非言語的情報のみ扱えるという前提の下での同期性に基づく協調系を観察するうえで，エージェントの個体数の議論は避けられない．2 体のエージェント間における同期性を考えるならば，個体数を増加させると同期的に振る舞う機会が増えるため，協調系を創り出すことが容易になる．しかしながら，缶蹴りの場合は子の数が増えるとゲームとしての難しさが変化するため，単純な比較はできない．なぜなら，子の数はダイナミクスに直接的に影響し，エージェント個体の視点では全く移動せずに報酬を獲得する，という状況もあり得てしまうからである．この問題を解決するためには，個体数を変化させてもゲームの難易度が変化しないようなルールを用意する必要があると考えられる．

6　おわりに

本研究では，競争と協調の両方の性質が含まれ，さらに共適応によって協調系が適応的な戦略では無くなる可能性のある缶蹴りを題材として，エージェントに強化学習させることでボトムアップに協調的な振る舞いを創発させた．今回は，同期性に基づいた戦略の下で構築される協調系に注目した．学習の結果，子は他者を直接知覚しなくても，統一された目標としての報酬のフィードバックによって協調的な振る舞いが可能であることが示唆された．

本研究の限界として，他者知覚の有無と協調可能性との関係への言及は，缶蹴りにおける鬼のように，全ての子と関わる可能性のある中心的なエージェントが存在する環境のみに限られる．缶蹴りにおける鬼と子の関係と同様の，中心的なエージェントを介したインタラクションによる集団構造が見られる環境は，例えばオーケストラにおける指揮者と演奏者の関係や，免疫システムにおける細菌と白血球の関係等が挙げられる．また，移動の同期性という振る舞いのレベルのみに焦点を当てているため，言語が介入する等，振る舞い以外の要因によってインタラクションが変化する状況についての検証が必要である．

我々が日常的に行っている協調について理解するためには，缶蹴りのような，協調系が動的に変化する共適応的な環境の中で，どのように協調系を創り出していくか，協調するための必要条件は何か，ということを検証することが重要である．缶蹴りのような協調戦略が一意に定まらないマルチエージェント系に創発する協調系の分析から，人と協調可能な人工物の設計や．人同士の協調を促すコミュニケーションシステムの構築が期待される．

参考文献

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

[2] Reynolds, C. W.: Flocks, herds and schools: A distributed behavioral model, In Proceedings of the 14th annual conference on Computer graphics and interactive techniques, pp. 25-34 (1987)

[3] 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)

[4] 中嶋宏，森島泰則，山田亮太，川路茂保：人間–機械協調システムにおける社会的知性 -心のモデルとパーソナリティによるエージェントの社会的応答について-，人工知能学会論文誌，Vol. 19，No. 3，pp. 184-196 (2004)

[5] Sen, S. and Sekaran, M.: Multiagent coordination with learning classifier systems, In International Joint Conference on Artificial Intelligence, pp. 218-233 (1995)

[6] 荒井幸代，宮崎和光，小林重信：マルチエージェント強化学習の方法論: Q-learning と profit sharing による接近，人工知能学会誌，Vol. 13，No. 4，pp. 609-618 (1998)

[7] 椿本樹矢，小林邦和：意図推定法を用いたマルチエージェント強化学習システムにおける協調行動の獲得，電気学会論文誌 C(電子・情報・システム部門誌)，Vol. 135，No. 1，pp. 117-122 (2015)

[8] Raileanu, R., Denton, E., Szlam, A., and Fergus, R.: Modeling others using oneself in multi-agent reinforcement learning, arXiv preprint arXiv:1802.09640 (2018)

[9] 大倉和博，保田俊行，松村嘉之：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（C 編），Vol. 77，No. 775，pp. 399-412 (2011)

[10] 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)

[11] Uwano, F. and Takadama, K.: Reward Value-Based Goal Selection for Agents' Cooperative Route Learning Without Communication in Reward and Goal Dynamism, SN Computer Science, Vol. 1, No. 3, pp. 1-18 (2020)

[12] Reynolds, C. W.: Competition, coevolution and the game of tag, In Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems, pp. 59-69 (1994)

[13] Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J. and Vicente, R.: Multiagent cooperation and competition with deep reinforcement learning, PloS one, Vol. 12, No. 4, e0172395 (2017)

[14] Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B. and Mordatch, I.: Emergent tool use from multi-agent autocurricula, arXiv preprint arXiv:1909.07528 (2019)

[15] Juliani, A., Berges, V. P., Vckay, E., Gao, Y., Henry, H., Mattar, M. and Lange, D.: Unity: A general platform for intelligent agents, arXiv preprint arXiv:1809.02627 (2018)

[16] 市川淳，藤井慶輔：協調に関する議論に向けたアプローチの提案: 集団運動からみる他者の行動予測と適応，認知科学，Vol. 27，No. 3，pp. 377-385 (2020)

[17] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O.: Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017)

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

[19] Dawkins, R. and Krebs, J. R.: Arms races between and within species, Proceedings of the Royal society of London, Vol. 205, pp. 489-511 (1979)