缶蹴りにおける強化学習を通した集団の組織化とダイナミクスの分析

Analysis of Group Organization and Dynamics through Reinforcement Learning in “Kick the Can” Game

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

静岡大学
Shizuoka University

Abstract: 実世界に生きる生物が，環境や他者の情報を部分的にしか観測できない中でも環境や他者の変化に適応し，目的を達成できるのはなぜだろうか．本研究では，部分観測している環境や他者との相互作用が，観測していない環境や他者の状態やその変化を間接的に伝達しているという仮説の下で，このような限定的な情報の中でのインタラクションと集団ダイナミクスを分析し概観することを目的とする．題材として，部分観測ゲームである缶蹴り遊びを用いた．強化学習によるシミュレーションの結果，部分的な観測しかできない中で，缶蹴りのルールに基づいて組織化された集団としての性質が創発した．この結果より，他者の状態を直接観測できなくても，他者によって引き起こされるゲームの結果が共有されることで，目的を達成する振る舞いに適応可能であることが考えられる．本研究により，缶蹴りは社会に立脚して生きる実世界志向のインタラクションとそのダイナミクスについて議論するためのプラットフォームとなり得ることが示唆された．

1 はじめに

ゲームはルールによって勝敗が明確に分かれるため実力の評価が容易であることから，人の知的な戦略やインタラクションを分析し理解するためのプラットフォームとして古くから研究の題材として用いられてきた．2016 年に “AlphaGo” が囲碁の世界チャンピオンに勝利した [1] ことを皮切りに，翌年には人の棋譜データを用いずにゼロから戦略を学習する “AlphaGo Zero” が “AlphaGo” の実力を上回り [2]，さらに “AlphaZero” は囲碁のみならず将棋やチェスにおいても既存の強豪 AI の実力を上回る [3] など，ゲームを題材にした研究は急速に進展してきた．近年ではボードゲームの枠を超え，より複雑であるリアルタイムストラテジー（RTS）ゲームにおいて “AlphaStar” がプロゲーマーを打倒し [4]，“MuZero” は囲碁・チェス・将棋に加え一般的なデジタルゲームにおいてルールの知識すら無い段階からルールと戦略を学習し強豪 AI の実力を上回った [5]．

以上のように，ゲームを題材にした研究は近年特に盛り上がりを見せている．しかしながらこれまで研究されてきたゲームの多くは，ゲーム内の現在・過去の状態や行動が全てのプレーヤに平等に開示される完全情報ゲームである．完全情報ゲームでは，プレーヤは完全観測（perfect monitoring）が可能であり，観測情報に不確定要素が無いため，ゲーム木などを用いて合理的な手の探索を行うことが有効に機能することがわかっている [6]．一方，他プレーヤの状態や行動に不確定要素が存在する不完全情報ゲームは未だ十分な研究が行われておらず，不完全情報ゲームでのコンピュータプレーヤの実力は発展途上である．

現実世界に目を向けると，我々の生きる社会においては他者や環境についての観測不可能な情報を前提にインタラクションや戦術が展開される．すなわち，現実世界は実質的には不完全情報ゲームにならざるを得ない．不完全情報ゲームでは，プレーヤは環境や他プレーヤの状態・行動を不完全観測（imperfect monitoring）しかできないため，不足する情報を推定することがゲームの攻略に対して有効な方法となる．そのため将棋や囲碁などの完全情報ゲームと比べて，ポーカーや人狼などの不完全情報ゲームは言語・非言語コミュニケーションが頻繁に行われる．このように他プレーヤとのコミュニケーションが戦略の構築において重要な要素となる不完全情報ゲームでは，プレイヤ間の意思疎通や協調を行うためのインタラクションを観察できる．以上を踏まえると，他者との意思疎通や協調といった実世界志向のインタラクションを考えるうえでは，不完

*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

全情報ゲームにおけるインタラクションをモデル化し，そのダイナミクスを分析する必要があろう．

不完全情報ゲームはポーカーなどのボードゲームに限定されるわけではない．ゲームに参加する各プレーヤが身体を持ち，各々の限定された視点から情報を収集し行動する場合，プレーヤによって観測する情報が異なるという点でそのゲームは不完全観測な不完全情報ゲームと言える．実世界志向の不完全情報ゲームについて考えるうえでは，ボードゲームより身体性に基づくゲームを題材にすることが望ましい．本研究では，身体性に基づいた不完全情報ゲームである缶蹴り遊びを題材としてエージェント間の協調的インタラクションを創発させ，そのダイナミクスを明らかにすることを目的とする．缶蹴りはエージェント毎に視点が異なり，障害物に隠れるという行動によって他者の状態が不確定となる不完全情報ゲームである．他者の状態が不確定な状況下では，他者の行動による環境の変化を学習し適応することで協調的インタラクションが現れる．そのため本研究では，生物の学習メカニズムを模倣した手法である強化学習を用いて協調的なインタラクションをモデル化する．本研究の成果は，現実世界という不完全な情報の中で生きる人や人工物のインタラクションと，それによって生じる集団ダイナミクスを議論するための指標を提示することに貢献し得る．

2 関連研究

不完全情報ゲームは不確定要素を含むという性質上，他プレーヤの状態を推定するためのインタラクションが生じることが多く，それによってダイナミクスが複雑になるため，長らく研究対象から遠ざけられてきた．しかし近年では，チェスや将棋，囲碁に次ぐプラットフォームとして，不完全情報ゲームを題材にした研究が盛んに進められている．

2019 年には 6 人対戦のポーカーにおいて，facebook とカーネギーメロン大学が開発した “Pluribus” が人のトッププロに勝利した [7]．同年，麻雀において Microsoft が開発した “Suphx” が天鳳 10 段を達成し，人のトッププレーヤに匹敵する実力を見せた [8]．以上に挙げたポーカーや麻雀といったボードゲームは，ゲームを行うためにプレイヤの身体が必須ではないため，他プレーヤの状態推定のために複雑な身体動作を考える必要が無く，確率的な方法が用いられていた．しかし実世界志向のインタラクションを考えると，人は身体性に基づいて他者の状態を推定し意思決定を行なっているため，身体動作によるゲームを考える必要がある．

プレーヤが身体を有し，身体運動によって進行するゲームとして，鬼ごっこ [9] やかくれんぼ [10] が題材とされてきた．これらのゲームは，エージェントの身体に立脚した視界を持たせることで，プレーヤ毎に観測する情報の差異が生じるため，不完全情報ゲームに分類される．鬼ごっこやかくれんぼでは，ゲームのルールによって “攻める” 役割と “守る” 役割が定められており，攻守関係が交代することは無い．実世界志向のインタラクションを考えると，連続時間の中で常に攻守関係が交代し得るゲームのダイナミクスを考える必要がある．

3 缶蹴り

3.1 缶蹴りの特徴

缶蹴りは伝統的な遊びのひとつで，身体性を伴う不完全情報ゲームである．単純なルールであるにも関わらず，鬼（Tagger）と子（Player）の駆け引きや，鬼同士・子同士の協調といった多様なインタラクションが観察できる．缶蹴りの特徴は，そのインタラクションの構造にある（図 1）．鬼と子の戦略を最もミクロに捉えると，鬼の戦略は “子を探しに行くか，缶の近くに留まるか” という選択となり，子の戦略は “缶を蹴りに行くか，壁の後ろに隠れるか” という選択となる．つまり，鬼と子の 1 対 1 の攻防が缶蹴りにおけるミニマムなインタラクションであり，これらが集積してダイナミクスを形成する．また，重要な点は鬼と子が互いに “攻める” 手段と “守る” 手段を有していることであり，これらの選択を連続時間の中で判断することで鬼と子の攻守関係が常に交代し得る．以上をまとめると，缶蹴りの特徴は以下のようになり，鬼ごっこやかくれんぼといった従来の課題より実世界に近い状況でのダイナミクスを観察することができる．

(1) 身体性に基づく不完全情報ゲームである
(2) 競争と協調が混在し同時に進行する
(3) 対立エージェントの攻守関係が常に交代し得る

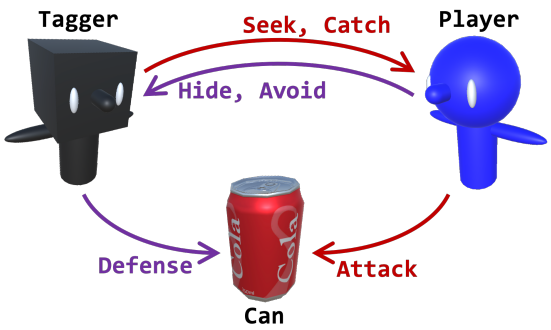

図 1: 缶蹴りのインタラクション構造

3.2 缶蹴りのルール

缶蹴りの手順とルールは以下の通りである．なお，本研究で焦点を当てない手順については省略する．

(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) 鬼が子を見つけた場合はシグナルを出し缶に触れることで，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．

3.2.1 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．子が能動的に行うことのできる戦略として，他の子が鬼に発見されずに缶を蹴るためには，他の子と同期的に缶に近づき，自ら鬼に発見される自己犠牲の振る舞いが想定できる．すなわち，“複数の子が同期的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．

3.3 缶蹴り環境

缶蹴りの環境は Unity[1]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 [11]．Unity で作成した缶蹴り環境を図 2 に示す．図 2 は，鬼 1 体と子 1 体のミニマムな缶蹴り環境における初期配置である．フィールド中央に缶があり，鬼は缶の周囲にランダムに配置される．円柱は缶から一定距離以上離れた位置でランダムに配置され，その後ろに子が配置される．なお，円柱はエージェントの視界を遮る障害物である．図 3 のように，エージェントは自身の身体位置に立脚した視界を有しており，前方 120 度に等間隔で 7 本の光線を飛ばし，光線に当たったオブジェクトの種類（鬼，子，缶，円柱，壁）と相対距離を認識するようにした．

[1]https://unity.com

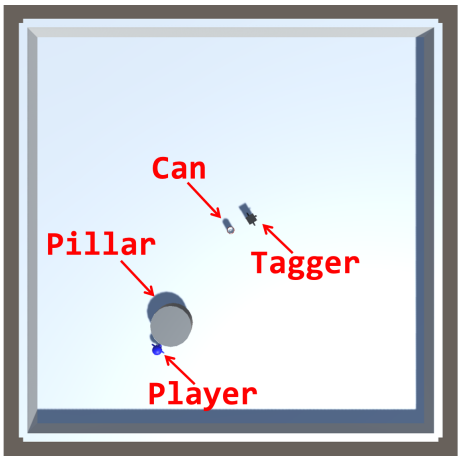

図 2: 缶蹴り環境（鬼 1 体 vs. 子 1 体の初期配置の例）

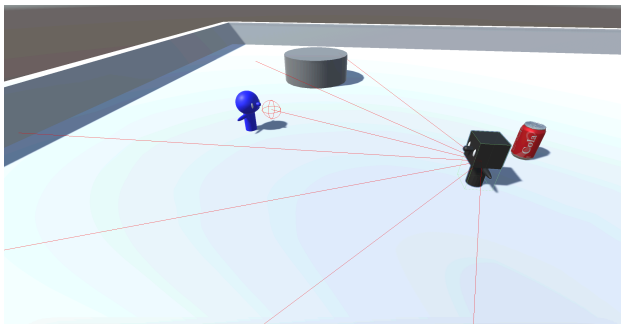

図 3: エージェントの視界と光線

4 強化学習

4.1 インタラクションの適応

あるゲームの状態で他プレーヤがどのように行動するか不明な状況において，他プレーヤとの競争や協調が求められる場合について考える．ルールベースなどにより意思決定則をトップダウンにモデル化する場合，全てのプレーヤの戦略は固定され，更新されない．その結果，ゲームのダイナミクスは初期戦略に依存してしまい，初期戦略で想定されなかった状況に対しては最適なインタラクションが行えない．そのため，できるだけ多くの状況について想定された初期戦略を考えなければならず，対象となるゲームについてのドメイン知識が必要となる [12]．

この問題を解決するためには，環境や他プレーヤとのインタラクションを通して意思決定則をボトムアップにモデル化する手法が有効となる．この手法では，環境や他プレーヤの状態に応じて戦略を更新することで，状況の変化に適応するインタラクションを記述できる．本研究では，エージェントの意思決定モデルをボトムアップに構築する手法として強化学習 [13] を用いた．強化学習は動物の行動決定則の変化をモデル化した手法

であり，その仕組みが動物の脳内に存在することを示唆する研究が行われている．Schultz et al. (1993) は，強化学習に用いられる報酬の期待誤差（TD 誤差）が脳神経におけるドーパミン反応と近似していることを明らかにし [14]，Barto (1995) や Schultz et al. (1997) によって，大脳基底核で TD 誤差を用いた強化学習が行われていることが示唆された [15][16]．さらに Doya (2002) は強化学習における割引率や学習率などのハイパーパラメータがそれぞれセロトニンやアセチルコリンなどの神経修飾物質と対応していることを提唱した [17]．以上のように，脳神経科学の観点から強化学習は人を含む動物の行動学習のモデルとして妥当であることが示唆されている．本研究では強化学習によって競争や協調のためのインタラクションの変化を実現することで，環境に立脚したプレーヤ間のインタラクションを観察した．

4.2 PPO

本研究では，ニューラルネットワークによって生物の学習プロセスを模倣した深層強化学習手法である PPO（Proximal Policy Optimization）[18] 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挙げている [18][19]．今回は，一般的な強化学習ライブラリである Unity ML-Agents[2] で PPO を使用した．図 4 に，PPO のフローチャートを示す．

PPO の特徴は，式 (1) を目的関数として勾配法を用いる点である．式 (1) 中の clip 関数によって，方策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される．clip 関数では，式 (2) に示す方策の変化量比が $1-\epsilon$ 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が行われる．なお，式 1 中の $\hat{A}_t$ は時点 t における Advantage（状態に依らない行動自体の価値）の推定値を表している．以上の処理によって，エージェントは自身の方策に対して急激な変化を行わないため，安定した学習が期待される．また，方策勾配は再帰型ニューラルネットワーク（Recurrent Neural Network, RNN）によって近似的に求められる．これにより，意思決定モデルの中で時間的な行動系列を扱うことができる．

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

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

[2] https://github.com/Unity-Technologies/ml-agents

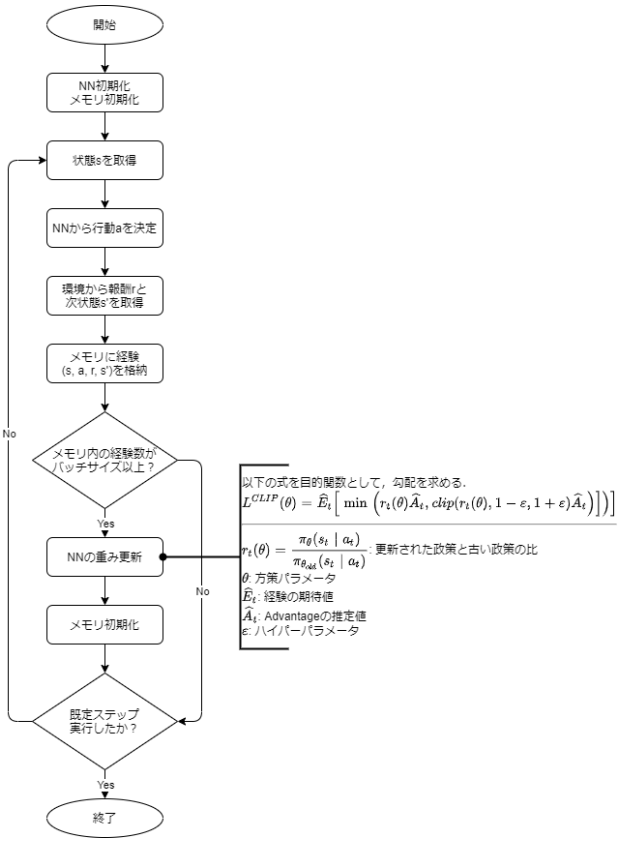

図 4: PPO のフローチャート

4.3 パラメータ

強化学習におけるエージェントの状態空間を表 1 に，行動空間を表 2 に示す．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 7 本の光線に当たっているオブジェクトの情報と，発見状態から構成される．発見状態は 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．

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	7
発見状態（鬼: 子を発見したか，子: 鬼に発見されたか）	1

表 2: 缶蹴り課題における行動空間

行動	値域
前後移動	[-3.0, 3.0]
左右移動	[-1.8, 1.8]
回転量	[-6.0, 6.0]

強化学習で使用する報酬を表 3 に示す．報酬は，3.2 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全ての子を捕獲することであり，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．そのため，これらの行動が達成された状態に対して報酬を設定した．

表 3: 缶蹴りの報酬

内容	報酬値		対象
	鬼	子	
子が缶を蹴る	-1	+1	集団
鬼が全ての子を捕獲する	+1	-1	集団
時間経過	-0.0001	+0.0001	個体

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．なお，今回の設定は Unity ML-Agents のデフォルト設定を用いた．

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	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

4.4 実行環境

本実験で用いたシミュレーション環境を表 5 に示す．

5 学習実験

5.1 実験パラメータ

缶蹴りなどのエージェントが分散的に意思決定と行動を行うシステムでは，エージェント数が 3 体以上になるとダイナミクスの予測が困難な複雑系としての多様なインタラクションが見られることが予想される [20]．そこで，缶蹴りにおいてエージェント数を増加させると，鬼集団・子集団のダイナミクスはどのように変化するか観察した．具体的には，鬼を 1 体から 5 体，子を 1 体から 5 体の範囲で変化させた 25 通りの組み合わせで実験を行なった．

表 5: 実験環境

種別	名称，バージョン，値
OS	Ubuntu 18.04.6 LTS
CPU	Intel Core i9-9980XE
GPU	GeForce RTX 2070
RAM	62.6 [GB]
シミュレーション環境	Unity 2020.3.10f
強化学習ライブラリ	Unity ML-Agents Release 18

5.2 結果と考察

集団サイズを変化させた場合における強化学習の結果得られた報酬の推移を図 5 に示す．図 5 では，表 3 における集団に対する報酬について示している．また，集団サイズを変化させた場合における，各エージェントの缶への接近量の変化を図 6 に，ゲームの最終状態の変化を図 7 に示す．なお，図 6 では各ステップあたり 10 回のトライアルを行い，その平均値をプロットしている．また，図 7 では各ステップあたり 10 回のトライアルを行い，ゲーム終了状態の回数を示している．

図 5 より，エージェント数の変化に対して集団ダイナミクスの変化が線形でないことがわかる．例えば，鬼が 1 体の場合に子の数を増やすと，4 体までは単調に子の獲得報酬が増加しているが，5 体になると 4 体より獲得報酬が低下している．これは，子が協調によって得点できる集団サイズに限界があることを示唆している．また鬼と子の数が同じ状況に注目すると，鬼 2 体/子 2 体のときは勝敗が明確に分かれて収束しているが，鬼 3 体/子 3 体では学習の途中から結果に揺らぎが生じており，これ以降は鬼/子の数が同じでも結果が収束しないことがわかる．このような非線形なダイナミクスの変化は，集団による創発現象と捉えることができる．

次に，図 6 より，エージェント数が増加するとエージェント毎に振る舞いのパターンが分化していることがわかる．例えば鬼 5 体/子 5 体の学習の最終ステップ付近では，子は缶に近づく振る舞いと近づかない（鬼から隠れる）振る舞いに分化しており，鬼は缶から離れない振る舞いと缶から離れる（子を探す）振る舞いに分化している．この結果から，個体としての振る舞いだけでなく，集団の中での組織化された役割を獲得したと言える．

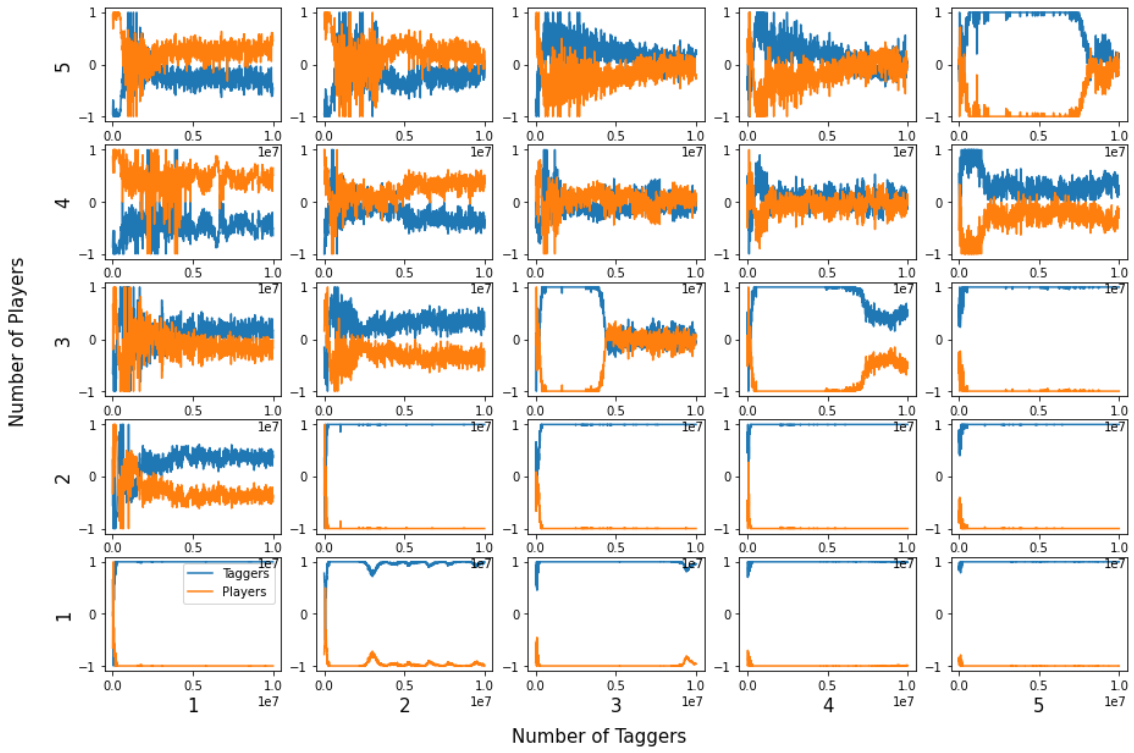

図 5: 缶蹴りの強化学習結果 エージェント数の変化に対する獲得報酬の推移

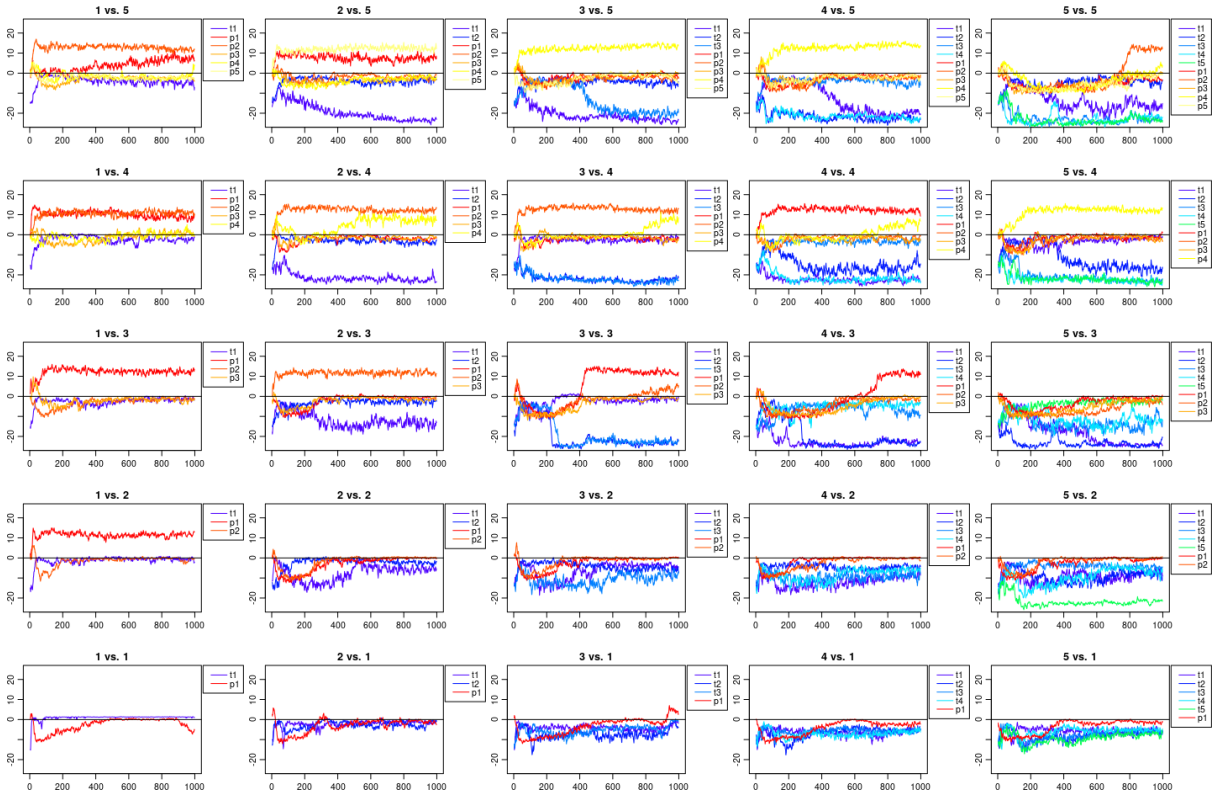

図 6: 学習結果 缶への接近量の変化

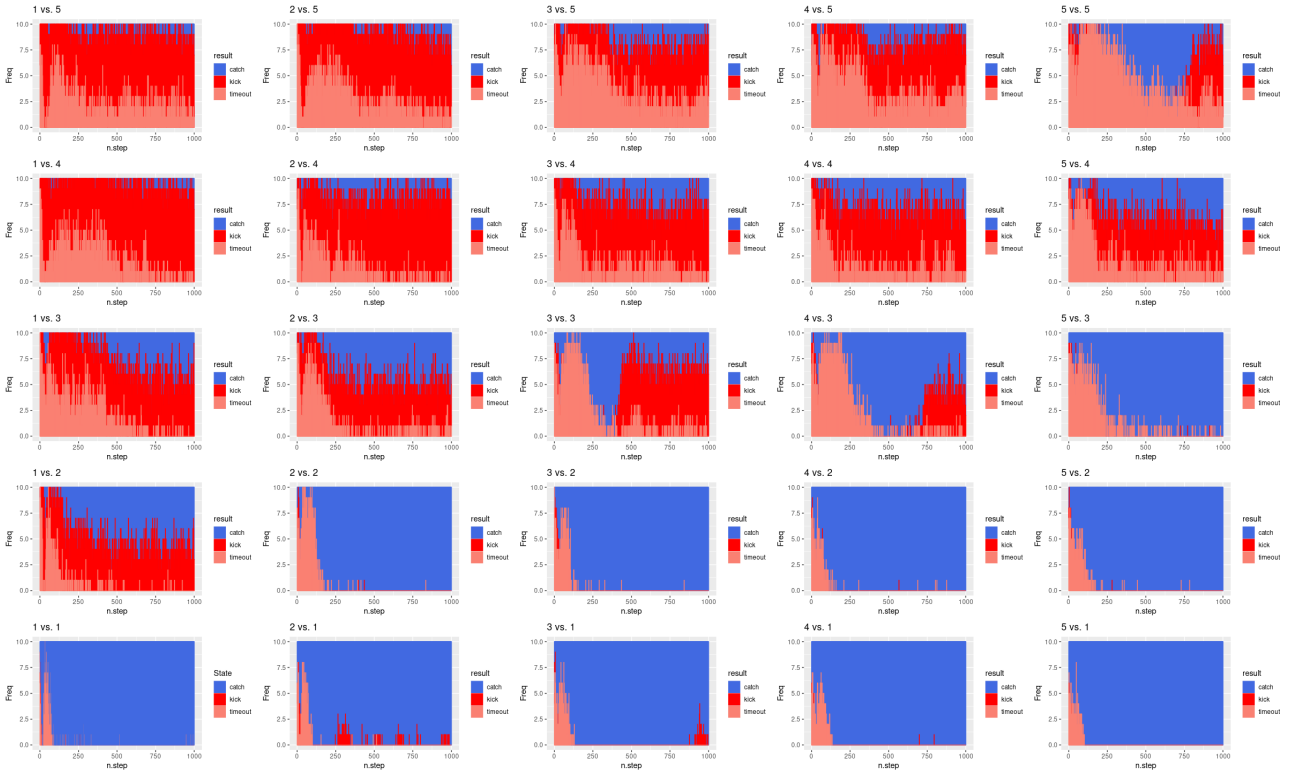

図 7: 学習結果 ゲーム終了状態の変化

また，図 7 より，鬼の数/子の数の増加によって，それぞれ表 6 のような変化の傾向があることがわかる．表 6 より，鬼の数に対する変化と子の数の増加に対する変化が対称ではないことがわかる．具体的には，子の勝利に関係する timeout 状態の回数が，子の数の増加に伴って増加する一方で，鬼の数の増加に対しては減少傾向ではないことが特徴的である．この結果は，子において timeout 状態によって勝利する戦略が鬼の増加に対して報酬を獲得しやすい戦略であることを示唆している．以上より，学習を通して，缶蹴り課題における子の戦略（kick 状態を目指すか，timeout 状態を目指すか）を鬼の数に対して適応的に獲得していると考えられる．

表 6: エージェント数とゲーム終了状態回数の変化傾向

エージェント数の変化	ゲーム終了状態回数の変化		
	鬼の勝利	子の勝利	
	catch	kick	timeout
鬼の数が増加	増加	減少	不定
子の数が増加	減少	増加	増加

6 議論

6.1 不完全情報ゲームとしての缶蹴りの協調系

缶蹴りでは記号的なコミュニケーションを行うことができない．そのため，他者と協調しようと思えば，環境や他者の行動の変化を注意深く観察し，その変化に適応するしかない．特に，子にとっては単体では鬼との勝負には勝てないため，鬼に勝利するならば子同士で協調的に振る舞う必要がある．このとき，缶蹴り課題においては，他の子が視界の外にいたり，障害物の後ろに隠れている場合，その状態や行動が不完全観測しかできない．

図 5 および図 6 に示した実験結果より，子は他者の身体を限定的にしか観測できないにもかかわらず，複数の子で缶に接近する協調的な振る舞いによって，集団の報酬を獲得できることが示唆された．このことは，他者の状態や行動が限定的にしか観測できなくても，他者の行動によって獲得される報酬が集団にフィードバックされ，子はその集団の報酬に応じて学習することによって協調系が創発したと推察される．他者が獲得する報酬は，子と缶または子と鬼とのインタラクションの結果のフィードバックと捉えることができる．すなわち，図 8 に示す構造が，缶蹴りにおいて協調系を創

発させることが可能であると考えられる．図 8 からもわかるように，缶蹴りでは缶や鬼が子の組織化を仲介している．このように，エージェント間を仲介するエージェントやルールが存在することで，間接的なインタラクションによる協調が可能であると考えられる．

この構造は，Peysakhovich et al. (2017) によって提唱された “結果主義的条件付き協力”（Consequentialist Conditional Cooperation, CCC）[21] と類似している．Peysakhovich et al. (2017) は，社会的ジレンマが生じる不完全情報ゲーム “Fishery” において，エージェントは過去の結果のフィードバックのみに注目することで協調できることを示した．Fishery は，異なるフィールドに存在する 2 体のエージェントが，フィールド間を行き来する魚を捕るゲームである．Fishery には競争や勝敗の概念は無く，相手が得点するために魚を捕獲せず残しておく協調的な戦略が求められる．一方，缶蹴りは同じフィールド内で競争と協調が同時に進行しているという点で Peysakhovich et al. (2017) の研究とは異なる．しかしながら，缶蹴りのような競争の中で協調が求められるゲームおいても，他者の観測は協調系の創発に必須ではなく，結果のフィードバックを観測することで協調可能，という CCC と同様の構造が確認された．

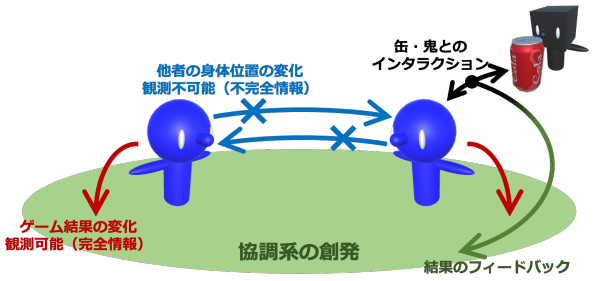

図 8: 缶蹴りにおいて協調系を創発させ得る構造

6.2 鬼と子による軍拡競争とゲームの揺らぎ

缶蹴りでは鬼と子の 1 対 1 の勝負では鬼が勝利するが，子同士の協調的インタラクションによって子が鬼に勝利できることがある．実験の結果（図 5）において，鬼 1 体/子 1 体などは鬼と子の獲得報酬が明確に分かれていて揺らぎがほとんど見られないが，鬼 4 体/子 4 体などは獲得報酬が 0 付近で揺らいでいることから鬼と子の勝敗が均衡状態であると考えられる．この均衡状態では，鬼と子の共適応による戦略の更新が頻繁に生じていると考えられる．すなわち，Dawkins et al. (1979) の提唱した軍拡競争 [22] の状態が継続していると言える．図 9 に，缶蹴りにおける鬼と子の軍拡競争の最も単純な循環を示す．実際には鬼や子は複数存在するため，それぞれの鬼や子によって異なる振る舞いの変化が生じている可能性があるが，この軍拡競争が継続されることによって勝敗の不安定さが継続していると考えられる．

Baker et al. (2019) のかくれんぼにおいても，鬼と子のインタラクションが共適応することで，子の戦略に鬼が適応し，その鬼の戦略に子が適応する，という結果が報告されている [10]．しかしかくれんぼの場合は，鬼と子がどちらも学習を続けると，最終的には鬼が侵入不可なシェルターを子が作り，その中に隠れることで子が勝利するという状況に収束している．このように，ゲームの戦略におけるひとつの解を見つけることはこれまで研究されてきた多くの題材で行われてきたが，今回得られた結果のようにゲームの勝敗が揺らぐ不安定な状況が安定して継続する結果は，缶蹴りに特有の結果である．缶蹴りは，かくれんぼや鬼ごっこなどの鬼と子が存在するゲームの中でも，鬼が子を捕まえることができるだけでなく，子が鬼を能動的に負かすことができるという特徴がある．このインタラクションのルールによって，図 9 のような軍拡競争のループ構造が生じ得ると考えられる．

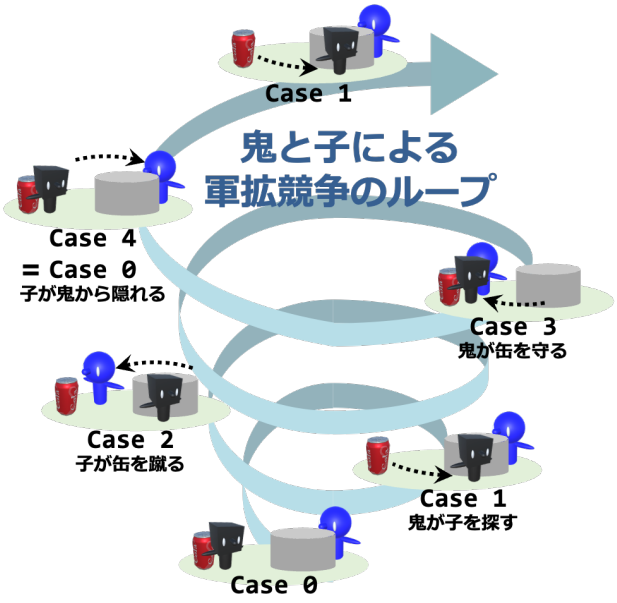

図 9: 缶蹴りにおける戦略の循環の例

6.3 ゲームの “面白さ” のパラメータデザイン

始めから勝敗が明確なゲームは面白さに欠け，勝負が拮抗していてプレーしてみるまで勝敗が不明なゲームが面白いように，一般的にゲームの “面白さ” には勝敗の不安定さによって評価可能な部分がある．6.2 節で

述べた缶蹴りの勝敗の不安定さは，集団サイズというパラメータを変化させた際に現れた性質であった．缶蹴りは軍拡競争がループすることでゲームの不安定状態が保持されていると考えられる．そのため，このような不安定なダイナミクスを保持するようにパラメータを調整すれば，図 5 における鬼 5 体/子 1 体のようなダイナミクスが収束する状態を避けることができ，ゲームの “面白さ” を維持することができると予想される．このように，ゲームのダイナミクスとパラメータの関係がわかることで，“面白い” ゲームを持続させるためのパラメータデザインが可能になると考えられる．本研究で行ったように，強化学習を用いてマルチエージェント系のダイナミクスを解析する手法は，ゲームの “面白さ” をデザインするための構成論的アプローチとして有効であると言える．

7　おわりに

本研究では，缶蹴り遊びを題材として，マルチエージェント強化学習によって集団の組織化が行われる様子を確認した．さらに，缶蹴りにおける子同士の協調系を創発させるために，缶や鬼と子とのインタラクションとそのフィードバックの観測による間接的なインタラクションが生じていることが示唆された．

本研究の限界として，身体的・認知的パラメータがエージェント毎に異なる場合について言及できない点が挙げられる．今回行った実験では，移動速度などの身体的パラメータやニューラルネットワークの構造などの認知的パラメータがエージェントに一様に設定されているため，個性パラメータに応じた集団の適応が行われない．実世界での集団は，必ずしも集団内のエージェントのパラメータが一様ではなく，個体の性質に応じた柔軟な協調系の様子が見られる．そのため，今後はエージェント毎に異なるパラメータ設定をして，個体の性質に適応した集団性を創発させることが課題となる．

Human-Agent Interaction においては，行為の帰結によって顕在化する他者性があるとされる [23]．本研究で観察されたエージェントもまた，その振る舞いに伴うゲームの帰結によって組織化された集団としての協調系が創発したと考えられ，Peysakhovich et al. (2017) の提唱した結果主義的な協調構造 [21] が缶蹴りの中に埋め込まれていることが示唆された．このように，缶蹴りを題材にすることで，エージェントが不完全観測の中で対立する目的を持つエージェントと相互適応しながら目的を達成するためのインタラクションを獲得する様子を観察し，分析することが可能であることが示された．本研究は，実質的に不完全情報ゲームである現実世界における，協調系が創発するためのインタラクションのデザイン，あるいは目的を達成するための組織化された集団を創発させる適応エージェントのデザインに寄与し得る．

参考文献

[1] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D.: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol. 529, No. 7587, pp. 484-489 (2016)

[2] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., & Hassabis, D.: Mastering the game of go without human knowledge, *Nature*, Vol. 550, No. 7676, pp. 354-359 (2017)

[3] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D.: A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, *Science*, Vol. 362, No. 6419, pp. 1140-1144 (2018)

[4] Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., H. Choi, D., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., P. Agapiou, J., Jaderberg, M., S. Vezhnevets, A., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., L. Paine, T., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps C., & Silver, D.: Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature*, Vol. 575, No. 7782, pp. 350-354 (2019)

[5] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D.: Mastering atari, go, chess and shogi by planning with a learned model, *Nature*, Vol. 588, No. 7839, pp. 604-609 (2020)

[6] Aumann, R. J.: Backward induction and common knowledge of rationality, *Games and Economic Behavior*, Vol. 8, No. 1, pp. 6-19 (1995)

[7] Brown, N., & Sandholm, T.: Superhuman AI for multiplayer poker, *Science*, Vol. 365, No. 6456, pp. 885-890 (2019)

[8] Li, J., Koyamada, S., Ye, Q., Liu, G., Wang, C., Yang, R., Zhao, L., Qin, T., Liu, T. Y., & Hon, H. W.: Suphx: Mastering mahjong with deep reinforcement learning, arXiv preprint arXiv:2003.13590 (2020)

[9] 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)

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

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

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

[13] Sutton, R. S., & Barto, A. G.: Reinforcement learning: An introduction, *MIT press* (2018)

[14] Schultz, W., Apicella, P., & Ljungberg, T.: Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task, *Journal of neuroscience*, Vol. 13, No. 3, pp. 900-913 (1993)

[15] Barto, A.G.: Adaptive critics and the basal ganglia, *Models of Information Processing in the Basal Ganglia*, pp. 215-232 (1994)

[16] Schultz, W., Dayan, P., & Montague, P. R.: A neural substrate of prediction and reward, *Science*, Vol. 275, No. 5306, pp. 1593-1599 (1997)

[17] Doya, K.: Metalearning and neuromodulation, *Neural networks*, Vol. 15, No. 4-6, pp. 495-506 (2002)

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

[19] Bøhn, E., Coates, E. M., Moe, S., & 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)

[20] Simmel, G.: The sociology of georg simmel, *Simon and Schuster*, Vol. 92892 (1950)

[21] Peysakhovich, A., & Lerer, A.: Consequentialist conditional cooperation in social dilemmas with imperfect information, arXiv preprint arXiv:1710.06975 (2017)

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

[23] 竹内勇剛: 原初的インタラクションからの他者の存在への気づき, 日本ロボット学会誌, Vol. 31, No. 9, pp. 850-853 (2013)