教示者と身体性が異なる学習者集団の模倣学習を通じた役割分担

Emergence of Social Roles through Imitation Learning between Groups with Different Embodiments

弓場　亮介 [1*]　堀井　隆斗 [2]　長井 隆行 [2,1]
Ryosuke Yuba[1]　Takato Horii[2]　Takayuki Nagai[2,1]

[1] 電気通信大学
[1] The University of Electro-Communication
[2] 大阪大学
[2] Osaka University

Abstract: 集団行動における子どもたちの行動獲得過程をモデル化することは，役割分担をはじめとする人の社会性発達の解明につながる．行動獲得には他者の行動を見まねする模倣学習が有用であることが知られているが，学習者と教示者で身体性が異なる場合（例えば子どもと大人）や集団内の学習者が異なる身体性を持つ場合の学習はあまり考慮されていない．本研究では実社会の子ども集団を想定し，教示者集団と学習者集団に異なる身体性を持つエージェントが存在する場合において，環境報酬を付与した模倣学習が役割分担行動創出につながることを示す．

1　はじめに

集団行動における子どもたちの行動獲得過程をモデル化することは，役割分担をはじめとする人の社会性発達過程を明らかにすることにつながる．人の社会性は幼児期に遊び場などで周囲の人との関わり合いを通じて獲得していくとされている．言葉が未発達な子どもたちにとって，身体を通した他者との関わり合いは重要である．子どもたちは言葉を介さずとも他者との関わり合いの中で行動を観察し見まねすることで行動を獲得できる．行動獲得を通じて社会性が発達していくと，集団内の一員として課題に取り組むことができるようになる．

人が集団で課題に取り組む際には，各個人が個別の役割を担い行動する．例えば，複数の子が鬼に捕まらないよう逃げる課題（鬼ごっこ課題）では，単に逃げ回るだけでなく囮役となり鬼をひきつけたり，物陰に隠れるといった行動によって鬼を狼狽させ逃げやすくする．人がこのような課題達成行動を獲得するためには，環境中で試行錯誤する強化学習や，既に行動を獲得している教示者の行動を見まねして行動を獲得する模倣学習が有用である．

近年，複数の学習者が模倣する状況を想定した模倣学習モデルが提案された [1][2]．これらの模倣学習モデルは，学習者が対応する教示者を見まね学習するため，

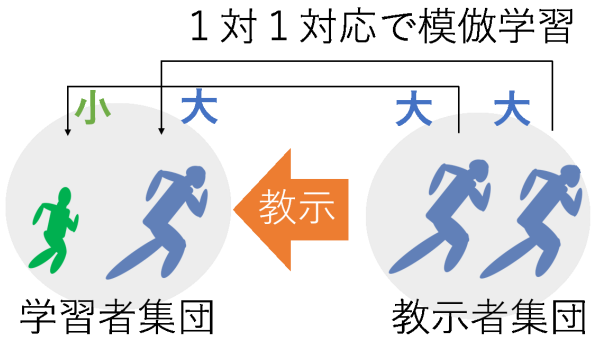

図 1: 異なる身体性を持つ学習者集団の模倣

教示者集団と同相な学習者集団による学習が想定されている．しかし，学習者と教示者で身体性が異なる場合（例えば子どもと大人）や集団内の学習者が異なる身体性を持つ場合の学習はあまり考慮されていない．例えば，保育園の遊び場で子どもたちが養育者を模倣する場面において，子どもたちは養育者と身体の大きさや足の速さなどの特徴が異なるが，各々が教示者を見まねし試行錯誤することで多様な行動を獲得できる．このような行動獲得過程のモデル化には，教示者と身体性が異なることを考慮し，集団内で自身が持つ身体性に合わせた振る舞いを獲得するための模倣学習モデル必要がある．学習者が教示者と異なる身体性を持つ場合，教示者の想定しない状態に陥ってしまうため，環境からの報酬を利用することで学習を効果的に進める

*連絡先：電気通信大学大学院情報理工学研究科
〒 182-8585 東京都調布市調布ケ丘 1-5-1
E-mail: r_aoki@apple.ee.uec.ac.jp

ことができると考えられる．

そこで，複数エージェントの敵対的逆強化学習 (Multi-Agent Adversarial Inverse Reinforcement Learning:MA-AIRL)[1] に課題達成の支援となる報酬 [3] を導入する．本研究では，課題補助報酬を付加した MA-AIRL による模倣学習を通じて多様な行動を獲得する手法を提案する．具体的には，逆強化学習によって推定された教示者報酬に課題達成の支援となる環境からの報酬を付加することで，学習者集団が教示者と異なる身体性を持つ条件での模倣学習の性能向上を図る．また，学習者集団が持つ身体性の組み合わせと，提案手法により役割分担の発現を目指す．

2 提案手法

2.1 学習者集団の特徴と行動方策獲得手法

学習者集団と教示者集団の身体性が異なる条件での模倣学習について説明する．本稿で仮定する模倣学習の集団構造を図 1 に示す．従来手法では対応する学習者と教示者の身体性は同一であるが，本研究では異なる身体性（例えば足の速さや手先の器用さが異なる状況）を仮定する．これにより最適方策に違いが生じる．模倣学習には MA-AIRL[1] を用いるが，教示者が想定しない行動の学習は困難であるため，黄瀬と谷口によって提案された手法 [3] を基に，環境からの報酬を敵対的学習の枠組みに導入することで多様な方策の獲得を目指す．また，異なる身体性から発現する個人固有の行動目的を表現するために，エージェント 1 体につき 1 つの識別器と 1 つの生成器を利用する．次節にて提案手法の基礎として，MA-AIRL について概説する．

2.2 MA-AIRL

Adversarial Inverse Reinforcement Learning (AIRL)[4] や Generative Adversarial Imitation Learning (GAIL)[5] は，単一エージェントが教示者行動から方策を学習するための模倣学習モデルである．AIRL や GAIL は最適な方策を持つ教示者の行動から教示者の報酬関数を推定する逆強化学習し，推定した報酬関数を用いて強化学習する．AIRL や GAIL は教師あり学習を基にした単純な模倣学習である Behavior Cloning (BC) よりも教示者方策を高い精度で復元できることを示した．また，教示者の行動と学習者の行動を区別する識別器と推定された報酬の期待積算を最大化するように方策を学習する生成器を持ち，それぞれ逆強化学習と強化学習に対応している．

MA-AIRL や MA-GAIL は模倣学習をマルチエージェント学習の枠組みに拡張したものである．これらのモデルは，マルコフ決定過程をマルチエージェントの枠組みに拡張したマルコフゲームの理論を基としている．また，複数の学習者がそれぞれ対応する教示者の報酬関数を推定し，推定した報酬関数に従って強化学習を行うことで方策を学習する．MA-AIRL は教示者の報酬関数を直接推定可能なため MA-GAIL よりも教示者方策を高い精度で復元できることがわかっている．

MA-AIRL の識別器はパラメータ ω に基づいて次式を解くことによってエージェントの報酬関数と状態価値関数を学習する．

$$\max_{\boldsymbol{\omega}} \mathbb{E}_{\boldsymbol{\pi}_E}\left[\sum_{i=1}^{N}\log\frac{\exp\left(f_{\omega_i}(\boldsymbol{s},\boldsymbol{a})\right)}{\exp\left(f_{\omega_i}(s,\boldsymbol{a})\right)+q_{\theta_i}\left(a_i|s\right)}\right]+ \mathbb{E}_{\boldsymbol{q\theta}}\left[\sum_{i=1}^{N}\log\frac{q_{\theta_i}\left(a_i|s\right)}{\exp\left(f_{\omega_i}(s,\boldsymbol{a})\right)+q_{\theta_i}\left(a_i|s\right)}\right] \tag{1}$$

$$f_{\omega_i,\phi_i}\left(s^t,a^t,s^{t+1}\right)=g_{\omega_i}\left(s^t,a^t\right)+\gamma h_{\phi_i}\left(s^{t+1}\right)-h_{\phi_i}\left(s^t\right)$$

ここで環境の状態を s，全てのエージェントの行動を $\boldsymbol{a}$，全教示者の方策を $\boldsymbol{\pi}_E$，全ての学習者の方策を $\boldsymbol{q}_\theta$ とする．また，エージェント i の報酬推定器は g_{ω_i}，状態価値推定器は h_{ϕ_i} である．

MA-AIRL の生成器はパラメータ θ に基づいて次式を解くことによってエージェントの方策を学習する．

$$\begin{aligned}&\max_{\theta}\mathbb{E}_{\boldsymbol{q}_\theta}\left[\sum_{i=1}^{N}\log\left(D_{\omega_i}(s,\boldsymbol{a})\right)-\log\left(1-D_{\omega_i}(s,\boldsymbol{a})\right)\right]\\&=\mathbb{E}_{\boldsymbol{q}_\theta}\left[\sum_{i=1}^{N}f_{\omega_i}(s,\boldsymbol{a})-\log\left(q_{\theta_i}\left(a_i|s\right)\right)\right]\end{aligned}$$

$$D_{\omega_i}(s,a)=\frac{\exp\left(f_{\omega_i}(s,\boldsymbol{a})\right)}{\exp\left(f_{\omega_i}(s,\boldsymbol{a})\right)+q_{\theta_i}\left(a_i|s\right)}$$

識別器と生成器が最適解を示す場合，f_ω は教示者方策のアドバンテージ関数を近似し，q_θ は教示者方策を近似する．

2.3 課題補助報酬を付加した MA-AIRL

本研究では MA-AIRL に課題補助報酬を導入することで，異なる身体性を持つエージェントの模倣学習において，環境からの課題補助報酬が学習者の行動獲得を支援することを目指す．MA-AIRL に課題補助報酬を導入したモデルを本稿では Multi-Agent Task Reward oriented Adversarial Reignforcement Learnig(MA-TRAIRL) と呼ぶ．MA-TRAIRL の枠組みを図 2 に示す．MA-TRAIRL において定義されるエージェント i の報酬関数は

$$\mathrm{R}_{\omega_i}(s,a)=\alpha g_{\omega_i}(s,a)+(1-\alpha)R_{task_i}(s,a)$$

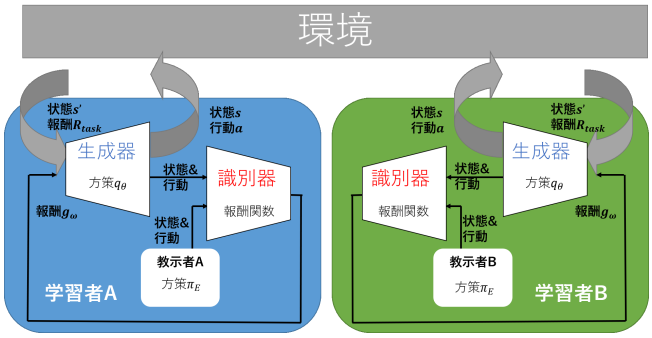

図 2: MA-TRAIRL の概念図

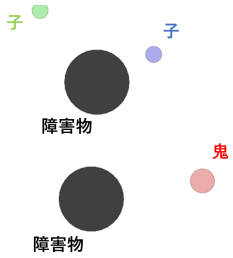

図 3: 実験環境（MAPE）

ここでエージェント i の課題補助報酬を R_{task_i} とする．重み α は報酬関数において模倣学習の報酬推定器が占める割合である．本稿では $\alpha = 0.5$ を採用した．また，MA-TRAIRL では生成器に Multi-agentActor-Critic with Kronecker-factors(MACK)[2] を用いる．MA-TRAIRL のアルゴリズムを Algorithm1 に示す．

本研究では教示者と異なる身体性を持つ学習者を想定しているため，学習者集団が教示者が想定しない状況に遭遇する．例えば，学習者の歩幅が教示者より小さい場合には，同じ位置から学習者が教示者と同じ方向に同じ歩数進んでも教示者と同じ位置にたどりつくことができない．MA-TRAIRL による集団の模倣学習では，教示者が想定しない状況で課題達成までに至らない場合，課題達成の補助となる環境報酬を得ることで，自分の状況を理解して課題達成のための行動を学習したり，他の学習者の状況を理解して課題達成のために助けることができると考えられる．また，学習者集団は，教示者集団と異なる身体性を持つため，各個人の最適方策に違いが生じて，多様な行動を獲得できると考えられる．

3 実験

3.1 実験目的

本実験では，教示者集団と異なる身体性を持つ学習者集団が教示者行動から模倣学習する際に，提案モデルが有用であることを検証する．また身体性の違いから獲得する行動特性が変化するかを検討する．本実験の目的は，既存の MA-AIRL モデルと課題補助報酬を導入した MA-TRAIRL モデルの性能を比較し，その妥当性を評価することである．MA-TRAIRL モデルを用いて学習することで，学習者は獲得する行動に違いが生じるかを検証する．

3.2 鬼ごっこ課題

本実験では Multi-Agent Particle Environment(MAPE)[6] を利用したマルチエージェント環境において，複数のエージェントが鬼から逃げる鬼ごっこ課題をモデル評価の対象とする．環境中には 1 体の鬼と 2 体の子，そして 2 つの障害物が存在する．図 3 に実験環境を俯瞰視した図を示す．この課題では鬼は子に接触することを，子は鬼との接触を避けることを目的としている．また環境中の障害物は進入不可領域となっている．各エージェントは毎時刻行動選択時に，上下左右への移動または静止の 5 種類の行動から 1 つを決定する．子の観測情報は，自分の座標と移動則と，他の子や鬼，それぞれの障害物との距離，全ての子の行動選択回数となっている．また今回の実験では鬼は最近傍の子へ向かって直線的に接近する行動を示す．

3.3 実験条件

模倣学習に利用するための教示者軌道は実験環境において上記に示したルールに基づいて行動する鬼と，鬼との接触によって負の報酬を与えられる 2 体の子を MACK による純粋な強化学習によって学習した方策から獲得した．獲得された教示者軌道の例を図 4 に示す．

学習者集団と教示者集団は青と緑色の子 2 体によって構成される．本稿では次の 2 種類の身体性を持つ集団を用いる．

- **集団 A:**青と緑の子の移動速度が同じで鬼よりも大きい
- **集団 B:**青の子の移動速度が鬼より大きく緑の子は小さい

教示者集団は集団 A とし，学習者集団は集団 A と B を想定する．集団 B は鬼より移動速度が小さい緑の子が含まれるため，鬼より移動速度が大きい子 2 体によって構成される集団 A よりも集団としての性能が低

Algorithm 1 MA-TRAIRL

Input: 教示者軌道データセット $\left\{\tau_j^E\right\}_{j=1}^M$（軌道数 M，教示者軌道 $\tau_j = \left\{\left(s_j^t, \boldsymbol{a}_j^t\right)\right\}$ は状態 s と行動 $\boldsymbol{a}$ の系列）；マルコフゲームのパラメータ (エージェント数 N, 状態空間 $\mathcal{S}$, 行動空間 $\mathcal{A}$, 初期状態分布η, 状態遷移確率 $\boldsymbol{P}$, 割引係数γ)

方策 q と報酬推定器 $\boldsymbol{g}$，状態価値推定器 $\boldsymbol{h}$ のパラメータ $\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\phi}$ を初期化

repeat

学習者方策 $\boldsymbol{\pi}$ から学習者軌道データセット $\left\{\tau_j^\pi\right\}$ をサンプリング:

$s^1 \sim \eta(s), \boldsymbol{a}^t \sim \boldsymbol{\pi}\left(\boldsymbol{a}^t|s^t\right), s^{t+1} \sim P\left(s^{t+1}|s^t, \boldsymbol{a}^t\right)$

$\left\{\tau_j^\pi\right\}$,$\left\{\tau_j^E\right\}$ から状態行動ペア $\mathcal{X}_\pi, \mathcal{X}_E$ をサンプリング:

for $i = 1, 2, ..., N$ **do**

式 1 においてパラメータ ω_i, ϕ_i 更新：

$$\mathbb{E}_{\mathcal{X}_E}\left[\log D\left(s, a_i, s'\right)\right] + \mathbb{E}_{\mathcal{X}_\pi}\left[\log\left(1 - D\left(s, a_i, s'\right)\right)\right]$$

end for

for $i = 1, 2, ..., N$ **do**

エージェント i の報酬関数を更新：

$$\mathrm{R}_{\omega_i}(s, a) = \alpha g_{\omega_i}(s, a) + (1 - \alpha) R_{task_i}(s, a)$$

報酬関数に基づいて方策パラメータ θ_i 更新

end for

until 収束

Output: 方策 $\boldsymbol{\pi}_\theta$ と報酬関数 R_{ω_i}

い．集団 A は青と緑の子がお互いの身体性を考慮して逃げ回る必要ないが，集団 B は移動速度が大きい青の子が鬼の標的となるよう誘導しつつ鬼から逃げなくてはならない．

実験は次のような条件で比較する．まず，従来想定されている教示者と学習者の身体性が同じ集団構造，すなわち教示者と学習者集団がどちらも集団 A の状況において，MA-AIRL と MA-TRAIRL を用いて模倣学習し比較する．同相の集団構造での模倣学習において MA-TRAIRL を用いることで，課題補助報酬が学習者に与える影響を検証する．次に，教示者と学習者の身体性が異なる集団構造，すなわち教示者は集団 A で学習者集団が集団 B で，MA-AIRL を用いて模倣学習を行い，従来の模倣構造との比較を行う．MA-AIRL は本研究で想定する集団構造のような学習者と教示者の身体性が異なる条件での学習を想定していないため学習困難であると考えられる．最後に，上記の集団構造において MA-TRAIRL を用いて模倣学習し MA-AIRL と比較する．課題補助報酬と身体性の違いから，学習者集団は課題補助報酬なしと比べて性能が高いかつ教示者と異なる行動を獲得できると考えられる．

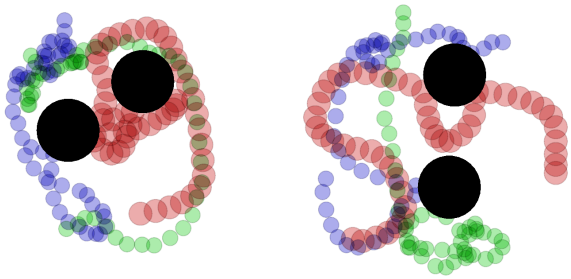

図 4: MACK による学習後の軌道（教示者軌道）左：緑の子が赤の鬼に追いかけられている様子．右：青の子が赤の鬼に追いかけられている様子

本実験では，MA-AIRL と MA-TRAIRL の事前学習として，教示者行動を BC で 500 イテレーション学習する．一般に GAIL 系統の模倣学習モデルでは，事前学習を行うほうが学習効率がよいといわれており，MA-AIRL[1] の学習でも BC による事前学習が行われているからである．MA-AIRL と MA-TRAIRL の学習は 50,000 イテレーション行う．

3.4 課題補助報酬：安全地帯

課題補助報酬は教示者が方策の学習に利用した報酬とは異なるが課題達成の支援となる報酬である．本稿では両方の子が安全地帯にいるか課題補助報酬を通じてを教示することで課題達成の支援とすることを試みる．

鬼からみて障害物の裏側にある領域を安全地帯と定義する．図 5 に示すように，鬼の位置によって安全地帯の領域は変化する．課題補助報酬 R_{task} は子 2 体両方が安全地帯にいない場合には負の値，そうでない場合には 0 となる．

鬼は子 2 体の位置を観測できるが，鬼は近傍にいる子に直線的に近づく方策をとる．鬼と子を挟むような位置に障害物があると，鬼が障害物と衝突するため，子は鬼との接触を回避できる．また，両方の子が安全地帯にいなければ負の報酬が与えられるため，一方の子が安全地帯にいる場合に，その子はもう一方の子が安全地帯にいるかどうか知ることができる．

3.5 評価指標

次に示す指標によってそれぞれの条件において学習された方策を評価する．今回の実験では方策の有効性と多様性の観点から評価指標を決定した．

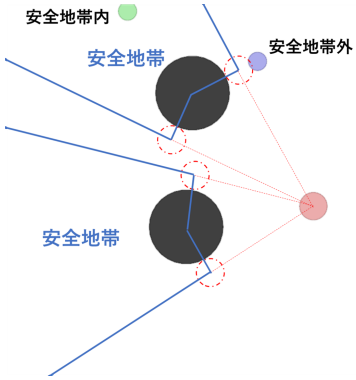

図 5: 安全地帯

鬼との接触回数

学習後のモデルを用いて，鬼ごっこ課題を 100 回実行し，鬼が子に接触した回数をカウントした．評価値は 100 試行当たりの接触回数平均値である．

相関係数

環境設定（子と鬼，障害物の位置）を固定した時の教示者軌道と学習者軌道のピアソンとスピアマンの相関係数を導出し，学習者方策や軌道の多様性を評価する．相関係数が小さい場合，教示者の行動方策とは異なる行動方策を学習者が獲得していると考えられる．100 試行それぞれで相関係数を算出し，その平均値を評価した．

3.6 実験結果

各条件での実験結果を表 1 に示す．まず，学習者集団が集団 A の構造のときの MA-AIRL と MA-TRAIRL の性能の違いを比較する．つまり補助報酬が学習に与える影響を評価する．それぞれのモデルによって学習された方策を利用して鬼ごっこ課題を実施した結果，MA-AIRL よりも MA-TRAIRL の方が鬼からの接触回数が少なくなった．これは課題補助報酬を模倣学習に組み合わせることで，方策の性能が向上したことを示している．また評価指標としての相関係数に注目すると，MA-AIRL よりも MA-TRAIRL の方が小さい値を示した．これは課題補助報酬の導入によって教示者の示した方策とは異なるものを獲得されたことを示している．

次に，学習者集団が集団 A と集団 B と異なる条件において模倣学習モデルに MA-AIRL を利用した際の性能を比較する．つまり集団が持つ身体性の違いが学習に与える影響を評価する．集団 A よりも集団 B の方が鬼からの接触回数が多くなった．これは集団 B は集団 A よりも集団としての性能が低いことや教示者と学習者で身体性が異なることが原因であると考えられる．相関係数に注目すると，緑の子は同じくらいの相関係数であるが，青の子は相関係数が小さい値を示した．青の子の相関係数に違いが生じたのは，一度緑の子が鬼に捕まると接触を続けてしまい集団としての模倣ができないため，青の子が試行錯誤したからだと考えられる．

集団 B のときの軌道例を図 6 を左図に示す．環境の初期状態において，緑の子は青の子よりも鬼との距離が近く鬼の標的となっている．緑の子は鬼との接触を繰り返し，青の子が鬼と緑の子の近くで近づこうとしたり遠ざかろうとしたりとうろうろとする様子が見られた．

最後に，学習者集団が集団 B のときの MA-AIRL と MA-TRAIRL の性能の違いを比較する．鬼からの接触回数に注目すると，MA-AIRL よりも MA-TRAIRL の方が少なくなった．これは集団 A のときの比較と同様に，課題補助報酬を模倣学習に組み合わせることで，方策の性能が向上したことを示している．また，課題補助報酬の導入により教示者との相関係数が小さい値をとったため，教示者が示した方策とは異なるものを獲得されたことを示している．

学習者集団が集団 B のときの MA-TRAIRL の軌道例を図 6 の右図に示す．MA-TRAIRL の学習軌道では，青の子は上方向へ進行することで鬼に接近し，鬼の標的を緑の子から自身へと誘導するように行動した．今回の学習に用いた教示者軌道では両方の子の移動速度が鬼よりも大きく，鬼から逃げ回る方策が表現されている．一方で，学習者集団が集団 B の場合は緑の子の移動速度が鬼よりも小さく，前述に実験結果（図 6 左）のように教示者行動を単純に模倣し逃げ回るだけでは鬼との接触を避けられない．MA-TRAIRL の学習軌道では課題補助報酬が影響し，両方の子が安全地帯へ入るような行動を示した．緑の子は鬼から遠ざかるように障害物へと近づき，青の子は鬼に近づき誘導してから安全地帯へと移動することで子 2 体が課題補助報酬を受け取れる．子 2 体が受け取る課題補助報酬が同じであるにも関わらず，緑の子と青の子は異なる行動特性を獲得した．身体性の違いと課題補助報酬により，緑の子は先に障害物へと隠れる役割，青の子は鬼を誘導する役割を担っているような行動を示した．

4 おわりに

本稿では，教示者と身体性が異なる学習者集団が MA-TRAIRL による模倣学習を通じて多様な行動を獲得する手法を提案した．実験では学習者と教示者の身体性が異なる場合には模倣学習が困難であるが環境から報酬を導入することによって学習される方策の性能が改善した．今後の課題は，MA-AIRL，MA-TRAIRL が

表 1: 100 回試行の平均接触回数と平均相関係数

手法	学習者集団	接触回数	ピアソン相関係数		スピアマン相関係数	
		青+緑	青	緑	青	緑
MA-AIRL	A	4.040	0.2799	0.3104	0.2591	0.3132
MA-TRAIRL	A	**1.670**	**0.1334**	**0.1089**	**0.1300**	**0.0910**
MA-AIRL	B	20.44	0.2465	0.3159	0.2076	0.2749
MA-TRAIRL	B	**11.96**	**0.2061**	**0.1685**	**0.2054**	**0.1191**

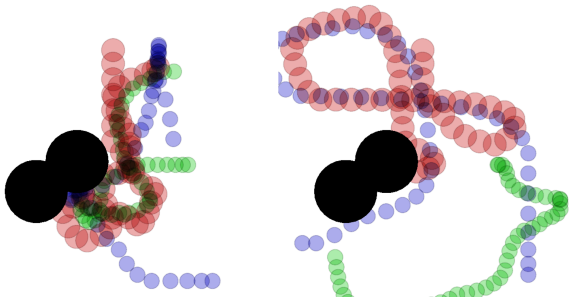

図 6: 左：集団 B-MA-AIRL，右：集団 B-MA-TRAIRL の学習軌道例．初期状態は緑の子の方が青の子よりも鬼との距離が近い．左図は緑の子が鬼との接触を繰り返している様子，右図は青の子が鬼に近づいてから逃げる様子が見られた．右図では鬼の標的が緑の子から青の子へと入れ替わりが見られた．これは課題補助報酬の導入によって，青の子が鬼を誘導し 2 体の子が安全地帯に入るように方策の獲得に影響したためである．軌道が欠けているのは可視化範囲をはみ出しているため．

それぞれの条件で獲得した方策をより詳細に解析するとともに，今回の条件とは異なる身体性を持つ学習者集団や課題補助関数に関して調査する．また，模倣報酬重み α を変更することによる獲得される方策の変化を検証する．

謝辞

本研究は，JST CREST (JPMJCR15E3)，新学術領域「認知的インタラクションデザイン学」の助成を受けたものである．ここに感謝の意を表する．

参考文献

[1] Yu, L., Song, J. & Ermon, S.. (2019). Multi-Agent Adversarial Inverse Reinforcement Learning. Proceedings of the 36th International Conference on Machine Learning, in PMLR 97:7194-7201

[2] Song, J., Ren, H., Sadigh, D., & Ermon, S. (2018). Multi-agent generative adversarial imitation learning. In Advances in Neural Information Processing Systems (pp. 7461-7472).

[3] 黄瀬輝, 谷口忠大. (2018). Generative Adversarial Imitation Learning にタスク達成報酬を付加した動作の学習. In 人工知能学会全国大会論文集 第 32 回全国大会 (2018) (pp. 2L2OS6a03-2L2OS6a03). 一般社団法人 人工知能学会.

[4] Fu, J., Luo, K., & Levine, S. (2017). Learning robust rewards with adversarial inverse reinforcement learning. arXiv preprint arXiv:1710.11248.

[5] Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. In Advances in neural information processing systems (pp. 4565-4573).

[6] Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In Advances in Neural Information Processing Systems (pp. 6379-6390).