マルチエージェント鬼ごっこ環境における深層強化学習エージェントと人の追いかけ行動の比較

Comparison of Chasing Behavior Between Deep Reinforcement Learning Agents and Human Participants in a Multi-Agent Predator-Prey Environment

西村 宏武，岡 夏樹*，田中 一晶
Nishimura Hirotake, Oka Natsuki, Tanaka Kazuaki

京都工芸繊維大学
Kyoto Institute of Technology
* nat@kit.ac.jp

概要

我々は人の社会的行動のメカニズムを構成的に解明することを目指している。本研究では、マルチエージェント鬼ごっこ環境を使用し、そこでの鬼側の深層強化学習エージェントの追いかけ動作と、同じ環境での人の追いかけ動作を比較した中間結果を報告する。移動エントロピーを指標として両者の間の相違点を検討したところ、興味深い違いが見つかった。今後はこの差異の原因を明らかにするため、再度動作比較実験を行う予定である。さらに、エージェントを人に近づけていくために、エージェントの設計仕様や差異の評価指標、個人差の原因について検討する計画である。

キーワード：社会的行動，深層強化学習，移動エントロピー，計算モデル

1. はじめに

認知モデルの立場からのマルチエージェントシステムの研究は古くから存在する[1-3]が、固定環境で単独で動作する場合と異なり、マルチエージェント環境における動作は、相手の動きとの相互作用で互いに動作が変化するため、複雑で予想や設計が難しく、これを人と比較して近づけていく研究は多くの課題を抱え、まだ発展の余地が大きいと考えられる。また、近年の深層学習の発展により、モデル化の手法や実現可能な機能が増えたが、ブラックボックス化し内部動作の解析が難しいという問題を抱えるようになった。

本研究では、マルチエージェント鬼ごっこ環境を使い、その中での追いかけ行動を深層強化学習でモデル化した。このエージェントの行動と、同環境でプレイした人の追いかけ行動を比較することで、モデルを人に近づけていくのが本研究の枠組みである。本論文では、この研究の第一報として、移動エントロピーを指標として両者の追いかけ行動を比較・分析した中間結果を報告する。

本論文の構成は以下の通りである。まず、本研究で使用したマルチエージェント鬼ごっこ環境（2 節）と深層強化学習エージェント（3 節）に関する情報を記し、続いて、エージェントと人の行動の比較・分析のための指標として用いた移動エントロピーについて簡単に紹介する（4 節）。5 節では、エージェントと人の行動の比較データの取得実験について述べ、6 節では実験結果を示す。最後に、この結果を考察し、今後の展望を記す（7 節）。

2. マルチエージェント鬼ごっこ環境

本研究では、[4]で使われた predator-prey 環境を若干改変したもの（図 1）を実験環境として採用した。

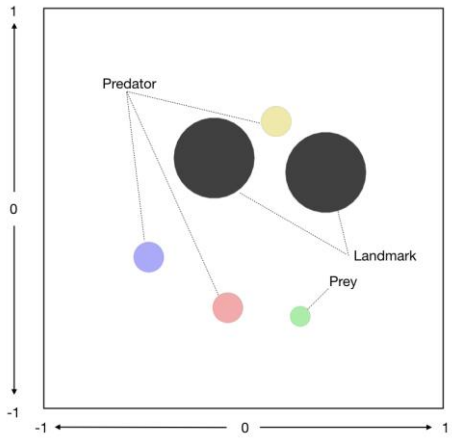

図 1　マルチエージェント鬼ごっこ環境

3 体の動きの遅い predator（鬼）が 1 体の動きの速い prey（子）を追いかける協力タスクであり、1 体の predator が prey に接触すると、3 体の predator が同じ大きさの正の報酬を得、prey が負の報酬（罰）を得る。2 個の landmark はランダムに配置され、障害物として働く。

3. 深層強化学習エージェント

本研究では次の 3 種類の深層強化学習エージェントを使用した。

- Deep Deterministic Policy Gradient (DDPG) [5]: deep Q-learning を連続行動空間に適用できるように改変したものである。
- Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [4]: DDPG をマルチエージェント向けに拡張したものである。学習時に他エージェントのポリシーを利用できることにして (centralized training with decentralized execution) 複雑なマルチエージェント環境での学習を可能にしている。
- Approximated Policy MADDPG (AP-MADDPG) [4]: MADDPG における、学習時に他エージェントのポリシーを利用できるという仮定を排除し、その代わりに、他エージェントのポリシーを推測することを学習し、それを用いたものである。

これら 3 種類のエージェントの追いかけ行動と人の追いかけ行動を比較して、どのエージェントが人のより良いモデルになっているかを調べ、人に近づけていくことを目指している（が、本論文の執筆時点では、まだその分析まで至っていない)。

4. 移動エントロピー

本論文では、他エージェントの行動 $Y(t)$ から自分の行動 $X(t)$ への移動エントロピー $T_{Y\to X}$ [6,7]を指標として、追いかけ行動の特徴を分析した。移動エントロピー $T_{Y\to X}$ は、他エージェントの行動 $Y(t)$ が自分の行動 $X(t)$ に関して持つ情報量であり、自分が他エージェントからどの程度の影響を受けて行動したかを表す。

本研究では、自分（行動 X の主体）が predator（鬼）の一員であるとして、自分が人である場合とエージェントである場合の移動エントロピー $T_{Y\to X}$ を比較することによりエージェントが人の追いかけ行動の特徴を捉えているかを調べることを試みた。他エージェント（行動 Y の主体）については、他エージェントが仲間の predator（鬼）である場合と prey（子）である場合の両方について移動エントロピー $T_{Y\to X}$ を調べる。

5. エージェントと人の行動の比較実験

鬼ごっこのすべてのプレイヤがエージェントの場合と、predator（鬼）の一人を実験参加者に置き換えた場合を比較するためのデータを取得した。

すべてのプレイヤがエージェントの場合については、predator（鬼）エージェントは、均一でない方が人同士の協調場面に近いと考え、3 節で説明した DDPG, MADDPG, AP-MADDPG 各 1 体ずつとし、prey（子）エージェントは基本的な機能を備えた DDPG とした。

ここで、predator（鬼）の観測情報は次の通りとした：

- 自分自身の絶対位置と絶対速度、
- landmark（障害物）の相対位置、
- 他の predator の相対位置
- prey の相対位置と絶対速度

なお、7 節で後述するように、他の predator の速度を観測情報に入れていないことが、人のモデルとしてはおそらく不適切であったことが実験結果の比較分析結果から示唆された。なお、上記の観測情報は、エージェントが行動を決める（すなわち、実験データを取得する）ときに利用する観測情報であり、MADDPG と AP-MADDPG の学習時には、他のエージェントの観測情報を利用するため、他のエージェントの速度も学習時に限り利用されていることを付記しておく。DDPG は学習時の観測情報も上記と同じである。

prey（子）の観測情報は次の通りとした：

- 自分自身の絶対位置と絶対速度、
- landmark（障害物）の相対位置、
- predator の相対位置

ここでも predator の速度を観測情報に入れておらず、人のモデルとしては恐らく不適切であると考えられるが、今回は prey を人とした場合とエージェントとした場合の比較は行わなかったため、結果とその考察に大きな影響はなかっただろうと考えている。

各学習エージェントは十分な学習を行った後で、学習を止めてデータ取得実験にて使用した。なお、[4]によると、predator, prey がすべて DDPG である場合と比べて、predator が MADDPG、prey が DDPG である場合には、エピソード当たりの接触回数（捕まった回数）が増加したことが報告されているが、今回の設定では大きな差はなかった。この違いの原因は、今後検討する必要がある。上記のように観測情報に他エージェントの速度を含まないことがこの違いと関係している可能性がある。

実験参加者は 7 名で、以下の手続きのデータ取得実験を行った。実験参加者は図 1 の画面が表示されたディスプレイを見ながら、椅子に座った状態で、ゲーム

コントローラを使用して、predator（鬼）のうちの1体を操作した。

データ取得実験の手続き
1. 操作説明と操作練習（4分程度）
2. 1体目のpredatorを操作して6分間プレイ
3. 2体目のpredatorを操作して6分間プレイ
4. 3体目のpredatorを操作して6分間プレイ

predatorエージェントは、DDPG, MADDPG, AP-MADDPG各1体ずつであったが、そのうちの1体を実験参加者が操作し、残りの2体のpredatorは学習は止めて自動で動作させた。上記のように実験参加者が操作するエージェントを順次変えて3回のデータ取得を実施した。操作するエージェントの順番は、実験参加者ごとに変えた。

以上の手続きで取得した、人が加わってプレイしたときのデータを、すべてがエージェントである場合（predator（鬼）エージェントは、DDPG, MADDPG, AP-MADDPG各1体ずつとし、prey（子）エージェントはDDPGとした）のプレイデータと比較した。

6. 実験結果

まず、すべてがエージェントである場合（predator（鬼）エージェントは、DDPG, MADDPG, AP-MADDPG各1体ずつとし、prey（子）エージェントはDDPGとした）において、他エージェントの行動 $Y(t)$ から自分の行動 $X(t)$ への移動エントロピー $T_{Y\to X}$ を30ステップ毎に算出した。他エージェントとしてはそれがpredator（鬼）であるかprey（子）を区別して分析するのでそれぞれを鬼、子と表記する。自分はDDPG, MADDPG, AP-MADDPGの3通りを区別して分析するのでそれぞれをD, M, Aと表記する。こうすると移動エントロピーとして、$T_{鬼\to D}, T_{鬼\to M}, T_{鬼\to A}, T_{子\to D}, T_{子\to M}, T_{子\to A}$ を区別することになる。

以下では、30ステップ毎に算出した移動エントロピーの分布の平均と歪度を散布図にプロットしたものを示す。

図2から図4は、それぞれ自分がDDPG, MADDPG, AP-MADDPGである場合の、他のpredator（鬼）からの影響の強さ（緑）とprey（子）からの影響の強さ（赤）の分布を平均（横軸）と歪度（縦軸）の散布図にしたものである。DDPG, MADDPG, AP-MADDPGの間には散布図上の分布の分散の大きさに差がある可能性があるが、その他に一言で表現できるようなはっきりした違いはなさそうである。しかし、他のpredator（鬼）からの影響の強さ（緑）とprey（子）からの影響の強さ（赤）の分布が（概ね）線形分離できるという共通の特徴があることが分かる。図2から図4を重ね合わせ、自分がDDPG, MADDPG, AP-MADDPGのいずれかで色分けし直した散布図が図5である。図5を見ると、分布が2つのクラスタ（他のpredator（鬼）からの影響の強さのプロットからなるクラスタとprey（子）からの影響の強さのプロットからなるクラスタ）にほぼ線形分離できること、DDPG, MADDPG, AP-MADDPGの間には明確な違いは無さそうなことが確認できる。

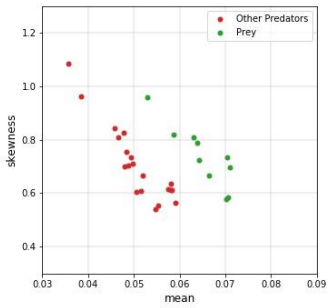

図2 $T_{鬼\to D}$（赤）と$T_{子\to D}$（緑）の散布図

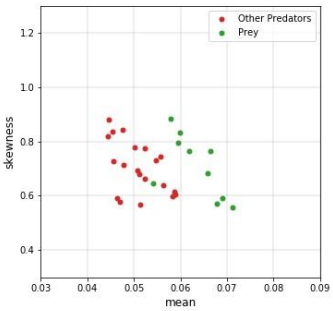

図3 $T_{鬼\to M}$（赤）と$T_{子\to M}$（緑）の散布図

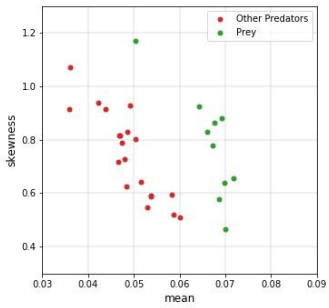

図4 $T_{鬼\to A}$（赤）と$T_{子\to A}$（緑）の散布図

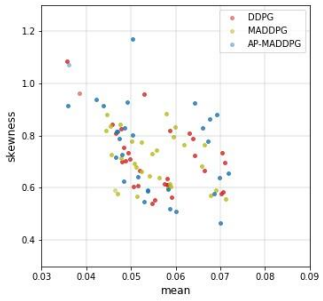

図 5　図 2 から図 4 の散布図の重ね合わせ

つづいて、predator（鬼）のうちの 1 体を人が操作した場合の $T_{鬼\to人}$ と $T_{子\to人}$ を図 5 に重ね合わせたものを図 6 から図 9 に示す。それぞれ実験参加者 1 から 4 までの結果である。$T_{鬼\to人}$ の分布と $T_{子\to人}$ の分布の間には明確な差はなかったため、両方を黒点でプロットした。ほぼすべての黒点が、図 2 から図 4 までの緑点が構成するクラスタ（prey（子）からの影響の強さを示す点からなるクラスタ）内[1]に存在することが分かる。残りの 3 名もこれら 4 人と同じ傾向であったため、結果の掲載は省略する。

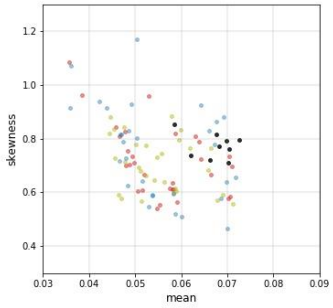

図 6　実験参加者 1 の移動エントロピー分布の図 5 への重ね合わせ

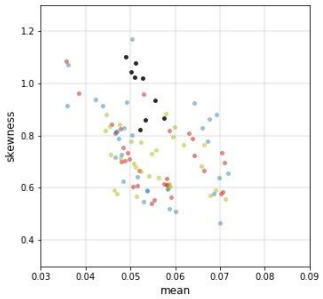

図 7　実験参加者 2 の移動エントロピー分布の図 5 への重ね合わせ

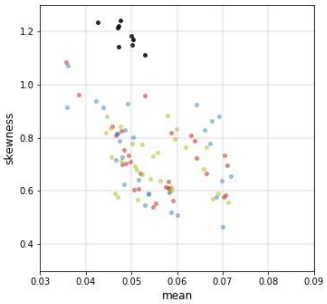

図 8　実験参加者 3 の移動エントロピー分布の図 5 への重ね合わせ

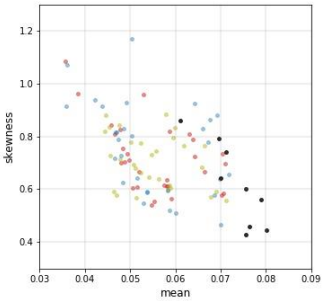

図 9　実験参加者 4 の移動エントロピー分布の図 5 への重ね合わせ

7.　考察と今後の展望

6 節の結果から、本研究で使用した強化学習エージェントと人の主要な違いは、他の predator（鬼）の動きからの影響の受け方であることが分かる。すなわち、人は predator（鬼）の動きからも prey（子）の動きからも同様な影響を受けているが、エージェントは prey（子）の動きからの影響の受け方は人と同様であるが、

[1] より正確には、2 クラスタを線形分離したと仮定した場合、prey（子）からの影響の強さを示す点が存在する半平面内

predator（鬼）の動きからの影響の受け方が人と異なる。

この違いの原因の候補（仮説）として現時点では以下の①～③を考えている。

① エージェントはprey（子）の速度は観測しているが、他の predator（鬼）の速度は観測情報に入っていないこと：
人の追いかけ行動はおそらく他の predator の現在位置だけでなく速度の影響も受けると思われる。これが正しいとすると、エージェントと人では、他の predator の速度を観測しているかどうかが異なることになる。一方で、prey についてはエージェントも人も共に速度を観測することになる。このことと、実験結果（散布図上で $T_{鬼→エージェント}$ だけが別のクラスタを構成する）は整合すると考えられる。つまり、「位置と速度を観測する他者から受ける影響と、位置だけを観測する他者から受ける影響は異なる」という仮説である。
この仮説の検証のためには、エージェントが他の predator の速度を観測するように仕様変更して再実験すればよい。本研究としてはこれが最優先の課題であると考えている。なお、エージェントの仕様をこのように変更すると、エージェントの動作が変わるためその影響で人の動作も変わることが予想される。このため、すべてのプレイヤがエージェントである場合の動作実験だけでなく、人を加えた実験も再度実施する必要がある。

② predator（鬼）と prey（子）が接触したとき受け取る報酬が人とエージェントで異なる可能性：
predator エージェントは、自分が prey と接触したか、他の predator が prey と接触したかに関わらず同じ大きさの報酬を受け取る。実験参加者にもそのように説明したが、実験参加者の気持ちとしては、自分だけで prey を捕まえた時よりも、他のエージェントと協調して捕まえた方が、よりうれしいといったことがあったかもしれない。その結果として、人は他のエージェントの動きも prey の動きと同程度に注目するということが起こり、それがクラスタ構成の差異につながった可能性がある。
報酬が異なることが原因という仮説の検証については、自分が捕まえたときと他の predator が捕まえたときで報酬の値を変えるという変更をするということであれば簡単だが、協調して捕まえた方がうれしいという変更の場合は、協調した程度を表す指標を新たに考案する必要があり、簡単ではないだろう。

③ 人とエージェントで学習の程度が異なった可能性：
エージェントの場合は、十分な学習をした後に、学習を止めてデータ取得セッションを実施したが、実験参加者の場合は、5 節に記した通り、4 分程度の説明と練習の後に、6 分間のデータ取得を 3 回実施した。このため、実験参加者は、データ取得前には十分な学習ができておらず、データ取得中にも学習が進んだ可能性がある。この場合は、人とエージェントのクラスタ構成の違いが、学習の進度の違いにより生じた可能性があると言えよう。人の結果は学習が未熟なことを反映しているという仮説である。
逆に、人はすぐに学習できるが、エージェントの学習が実際は不十分であった可能性もある。人の学習が未熟だという仮説の検証は、今回取得したデータを前半後半に分けて再分析するだけでも見通しがつく可能性があるため、まずこれを試みる計画である。4 分＋6 分×3 では全く不十分であった場合は、再実験が必要となる。エージェントの学習が実は不十分であった場合は、さらに学習を続けるか、または、学習方式やハイパーパラメータ等の再検討が必要な可能性もある。

上記①～③に加えて、以下の点も今後検討したいと考えている。

今回の実験では、他エージェントからの影響を受ける程度は、実験参加者ごとに大きい個人差があった（図 6 から図 9 を参照）。この個人差を説明できるモデルの構築は今後の課題である。

また、移動エントロピー以外の指標（たとえば、獲得報酬、学習速度など）による差異の分析も今後の課題である。

今回は、DDPG, MADDPG, AP-MADDPG の 3 種類の強化学習エージェントを使用したが、この選択の妥当性や他の学習モデルの検討も今後進めていきたい。

さらに、鬼ごっこ環境で得られた知見が、どこまで一般的に成立するかを、鬼ごっこ以外の課題における学習エージェントの振舞いと人の振舞いの比較を通して調べる必要がある。

文献

[1] Sun, R. (2001) “Cognitive science meets multi-agent systems: A prolegomenon,” Philosophical Psychology, Vol. 14, No. 1, pp. 5-28.
[2] Sun, R. (ed.) (2005) Cognition and multi-agent interaction: From cognitive modeling to social simulation, Cambridge University Press.
[3] Lenk, J.C., Droste, R., Sobiech, C., Lüdtke, A., & Hahn, A. (2012). “Towards cooperative cognitive models in multi-agent systems,” COGNITIVE 2012: The Fourth International Conference on Advanced Cognitive Technologies and Applications, pp. 67-70.
[4] Lowe, R.., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017) “Multi-agent actor-critic for mixed cooperative-competitive environments,” NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6382–6393.
[5] Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N.M., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015) “Continuous control with deep reinforcement learning,” arXiv:1509.02971.
[6] Schreiber, T. (2000) “Measuring information transfer,” Phys. Rev. Lett. Vol. 85, pp. 461-464.
[7] Lizier, J.T. (2014) “JIDT: An information-theoretic toolkit for studying the dynamics of complex systems,” Front. Robot. AI, Vol. 1, Article 11, 20 pages.