The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

人間の環世界から見たエージェントモデル –AI Safetyの実現に向けて

Modeling Agents from the Perspective of Human Umwelt towards AI Safety

福地庸介 *1 Yosuke Fukuchi　大澤正彦 *1*2 Masahiko Osawa　山川宏 *3*4 Hiroshi Yamakawa　今井倫太 *1 Michita Imai

*1慶應義塾大学理工学研究科 Graduate School of Science and Technology, Keio University

*2日本学術振興会 特別研究員 (DC1) Research Fellow of Japan Society for the Promotion of Science (DC1)

*3株式会社ドワンゴ ドワンゴ人工知能研究所 DWANGO Co., ltd Dwango Artificial Intelligence Laboratory

*4全脳アーキテクチャ・イニシアティブ The Whole Brain Architecture Initiative

Transparency in machine learning (ML) agents' decision making is crucial for achieving AI Safety. However, it is difficult to comprehend agents' behavior because gaps of perception, mobility, desire, and time scale between humans and ML agents obstruct people to mentalize the agents. In this paper, we propose a model of human's inference of ML agents' mental states so as to explain the agents' behavior from the perspective of humans.

1. はじめに

深層強化学習をはじめとする機械学習技術の進展により，ビッグデータと呼ばれる複雑で大量の情報をもとに行動を決定するエージェントが実現されている．またエージェントは，より複雑な行動を扱えるようになっている．

しかし，エージェントが複雑な情報処理を扱うようになるにつれて，エージェントの振舞いを理解することは難しくなっていく．エージェントの振舞いが理解不能となることは，人間と同じ環境を共にするエージェントにおいて特に深刻な問題である．人間の意図/予期しない振舞いが致命的な事故に繋がるためである．AI Safety[Amodei 16] を実現する上で，エージェントの振舞いに対する透明性を向上させ，人間がエージェントを理解できるようにすることが重要といえる．

機械学習によって獲得された振舞いの透明性を向上し，人間にエージェントを理解させる手法として，エージェントの行動の履歴から振舞いを確率的にモデル化して得られたモデルを自然言語で説明する方法が提案されている [Hayes 17]．また画像処理の分野では，大量のパラメータで表現されていて人間には理解が難しい深層学習器の分析が盛んに行われている [Le 13]．

いずれの手法も，行動決定を行う機械学習モデルが扱う情報空間の範疇で振舞いを説明している．しかしエージェントと人間は，同じ環境にあってもそれぞれ異なる観測空間の中で環境を知覚し，それぞれの時間・報酬・行動空間を持っている．本稿では，人間・エージェントそれぞれが固有に持っている情報空間を環世界と呼ぶ．人間とエージェント間の環世界の違いはエージェントの理解を難しくする．また人間が「盲導犬は信号の色を見て行動決定している」と誤解してしまうように，環世界の違いがエージェントを理解する際のバイアスとなり，エージェントの振舞いに対する致命的な誤解を生じさせる原因にもなる．

そこで本稿では，人間からはエージェントの振舞いはどのように理解されるのか，をモデル化した “人間の環世界から見たエージェントモデル” を提案する．人間の環世界から見たエージェントモデルの構築は，人間の情報空間の中でエージェントの振舞いを説明することや，人間がエージェントに対して抱く誤解を検知することにつながると考えられる．

連絡先: 慶應義塾大学理工学部情報工学科今井研究室
〒 223–0061 神奈川県横浜市港北区日吉 3–14–1
E-mail: fukuchi@ailab.ics.keio.ac.jp

2. 他者理解と環世界

2.1 人とエージェントの環世界

ユクスキュルは，生物が認識している世界が，生物種それぞれが持つ知覚と作用の上で構築される主観的な世界 (環世界) であると考えた [ユクスキュル 05]．環世界は，行動主体が持っている知覚器官からの情報で構成される知覚世界と，主体が働きかけられる対象から構成される作用世界の統一体とされる．

機械学習によって行動を獲得するエージェントもまた，行動によって環境に働きかけ，観測と報酬によって環境を知覚する，という一つの環世界を構築していると考えられる．

そこで本稿では，エージェント α の環世界 E^α を，エージェントが扱う状態空間 $\boldsymbol{S}^\alpha$，観測空間 $\boldsymbol{\Omega}^\alpha$，知覚時間 t^α と報酬関数 R^α からなる知覚世界，行動空間 $\boldsymbol{A}^\alpha$ からなる作用世界として，$E^\alpha = (\boldsymbol{S}^\alpha, \boldsymbol{\Omega}^\alpha, R^\alpha, \boldsymbol{A}^\alpha, t^\alpha, T^\alpha, O^\alpha)$ と考える．T^α, O^α はそれぞれ，エージェントが状態 $s^\alpha \in \boldsymbol{S}^\alpha$ のもとで行動 $a^\alpha \in \boldsymbol{A}^\alpha$ をとった際の状態遷移確率と観測確率である．同様に人間の環世界を $E^h = (\boldsymbol{S}^h, \boldsymbol{\Omega}^h, R^h, \boldsymbol{A}^h, t^h, T^h, O^h)$ とモデル化して考えることにする．

2.2 他者理解の壁となる環世界の違い

2.2.1 観測空間

人間の観測空間は，視覚や聴覚といった五感からの情報で構成される．一方磁力センサによって環境を認識するエージェントのように，人間が持っている観測空間と対応関係がない場合人間にはエージェントが環境の何を知覚できているのかを判断することが難しくなる．また，環境の状態の中でエージェントが知覚できていている範囲を推定することも難しい．

エージェントが環境の何をどのように観測しているかは，エージェントの振舞いを予測する際や，エージェントが環境に対して持つ信念を推定する際に重要な情報である．

2.2.2 行動空間

エージェントが関節角やモータの制御値のような低レイヤの行動を扱う場合や，蛇型エージェントのように人間の行動空間と離れた行動を扱うエージェントの場合，エージェントが扱う行動空間の値をそのまま人間が理解するのは難しい．

人間がエージェントの振舞いを理解し予測する際は，エージェントが実際に扱う低レイヤの行動空間ではなく，「右に行く」「手を伸ばす」といった，より抽象的な行動空間の中でエージェントの振舞いを捉えていると考えられる．

2.2.3 知覚時間

強化学習によって行動を獲得したエージェントの場合，エージェントの扱う時間ステップのたびに行動決定が行われる．しかしエージェントが行動を選択する際の時間間隔は，人間がエージェントを理解するには短すぎる場合が存在する．

例えば，深層強化学習を用いてゲーム内のエージェントを制御する [Mnih 15] では，エージェントが 4 フレームに一度行動を選択している．行動選択の時間間隔は 67 (msec) である．一方人間は，単一の刺激に対して単一の反応を行うだけでも 150 (msec) 以上の時間がかかるとされる [大山 85]．人間にはエージェントの知覚時間の単位の情報は扱いきれず，人間はより大きな時間粒度でエージェントの振舞いを捉えていると考えられる．

2.3 環世界の違いを超えた他者理解

人間とエージェントの環世界には大きなギャップが存在する．しかし機械学習によって行動獲得するエージェントの透明性の向上に関する従来研究では，エージェントの環世界の範疇でのみ振舞いが説明されてきた．環世界の乖離はエージェントの扱うタスクや行動が複雑になるにつれて大きくなると予想される．そのため人間が一方的にエージェントの環世界に合わせるのではなく，エージェントもまた人間の環世界を考慮して人間とインタラクションを行うべきだと考えられる．

マルチエージェントシステムの分野では，環世界の異なるエージェントがどのような信念や欲求を持っているかをエージェントの観測を元に推定する研究がある [Rabinowitz 18]．しかし Human-Agent Interaction の枠組の中で，機械学習によって獲得された振舞いが人間固有の環世界から見てどのように理解・解釈されるか，ということは考えられてこなかった．

そこで本稿では，人間が観測し理解できる情報だけの範疇でエージェントの振舞いを解釈し説明する「人間の環世界からみたエージェントモデル」を提案する．提案モデルでは，エージェントの観測空間と無関係に，人間の観測空間で得られる情報からエージェントの振舞いを扱う．またエージェントが選択する低レイヤーの行動ではなく，より抽象度の高いレベルでエージェントの行動を捉える．さらにエージェントが扱う時間粒度ではなく，より長期的な時間粒度でエージェントの振舞いをモデル化する．

環世界の違いから，提案するモデルはエージェントの真の行動決定と整合性の取れない場合が予想される．例えばエージェントが人間と異なる観測空間を扱う場合，人間の環世界からはエージェントの振舞いを説明しきれないと考えられる．提案モデルの目的は，エージェントの行動決定を正確に説明するモデルを構築することよりも，エージェントを観測する人間がエージェントに対して想定してしまう振舞いのモデル再現することにある．

2.4 扱う環境

本稿では，シミュレーション環境 [Catto] で機械学習による行動獲得を行なったエージェントを対象として，人間の環世界から見たエージェントモデルを考える．エージェントはロケット型でジェット噴射を操作することで特定の位置に移動する行動を獲得している．

エージェントの観測空間 $\mathbf{\Omega}^\alpha$ は，環境に関する 8 次元のパラメータ群で構成される．エージェントは固有の報酬関数 R^α を持ち，観測 ($\omega^\alpha \in \mathbf{\Omega}^\alpha$) と信念 ($b^\alpha \in \boldsymbol{B}^\alpha$) を元に行動 ($a^\alpha \in A^\alpha$) を選択する．

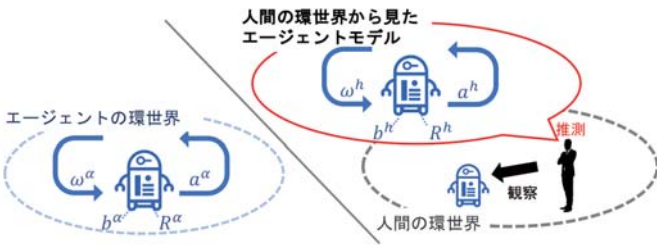

図 1: エージェント固有の環世界の中で行動決定するエージェント (左) と，人間の環世界から見たエージェントモデル (右)

3. 人間の環世界から見たエージェントモデル

3.1 モデル概要

人間の環世界から見たエージェントモデルは，エージェントの環世界 E^α と独立して，人間の環世界 E^h からエージェントを観察した際のエージェントの振舞いをモデル化したものである (図 1)．提案モデルは，[Baker 17] をもとに，エージェントの行動決定を E^h における部分観測マルコフ過程 [Kaelbling 98] として扱う．また提案モデルはエージェントが持つ報酬空間 R^α の情報は持たず，人間が想定するエージェントの信念状態と報酬関数の仮説の集合 $\boldsymbol{B}^h$，$\boldsymbol{R}^h$ を持ち，人間から見たエージェントの報酬関数 $R^h \in \boldsymbol{R}^h$ と信念 $b^h \in \boldsymbol{B}^h$ を確率分布として推定する．

3.2 実際のエージェントへの適用

本稿では，2.4 節で紹介したエージェントを対象に，人間の環世界から見たエージェントモデルを構築する．人間はエージェントを視覚により観察していると仮定して，$\mathbf{\Omega}^h$ はエージェントの位置情報と速度のみで構成されるとする．また人間が持つ時間粒度の大きさを考慮して，人間が扱う時間 t^h はエージェントが扱う時間ステップ t^α に対して $t^h = k \cdot t^\alpha$ とする．人間がエージェントに対して想定する行動空間 $\boldsymbol{A}^h$ を扱うため，つもり抽出 [丹羽 12] を参考にした事前実験を行う．事前実験では，人間がエージェントの動きを観察し、エージェントになったつもりになってゲームコントローラを操作させることで，人間がエージェントに対して想定する行動と，その際の状態遷移確率を抽出する．

4. おわりに

本稿では，AI Safety の実現に向けてた取り組みとして，人間からはエージェントの振舞いはどのように理解されるのか，をモデル化した “人間の環世界から見たエージェントモデル” を提案した．提案モデルは，人間の環世界の中でエージェントの振舞いを説明することや，人間がエージェントに対して抱く誤解を扱う際に重要なモデルであると考えられる．

今後は構築されるモデルを，実際に人間が考えるエージェントモデルと比較することで検証する．さらに，構築されるモデルにもとづいて，エージェントの振舞いを人間に理解させるためのインタラクションを検討していく．

参考文献

[Amodei 16] Amodei, D., et al. : Concrete problems in AI safety. arXiv preprint arXiv:1606.06565 (2016)

[Le 13] Le, Q. V. : Building high-level features using large scale unsupervised learning. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 85958598. (2013)

[ユクスキュル 05] ユクスキュル, クリサート (日高敏隆, 羽田節子訳), 生物から見た世界, 岩波書店. (2005)

[Hayes 17] Hayes, B., Shah, J. A. : Improving Robot Controller Transparency Through Autonomous Policy Explanation. Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. ACM, pp. 303312. (2017)

[Mnih 15] Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature 518, 7540 pp. 529533. (2015)

[大山 85] 大山正.: 反応時間研究の歴史と現状. 人間工学 21 (2) pp. 57-64. (1985)

[Kaelbling 98] Kaelbling, L. P., et al.: Planning and acting in partially observable stochastic domains. Artificial Intelligence 101, 99134 (1998).

[Catto] Catto, E. Box2d: A 2d physics engine for games. Accessed: 2017/5/26.

[丹羽 12] 丹羽真隆ら.: つもり制御: 人間の行動意図の検出と伝送によるロボット操縦 (<特集> テレイグジスタンスのためのロボティクス・グラフィクス・インタフェース). 日本バーチャルリアリティ学会論文誌 17.1. (2012)

[Rabinowitz 18] Rabinowitz, N. C., et al. : Machine Theory of Mind. arXiv preprint arXiv:1802:07740 (2018).

[Baker 17] Baker, C. L., et al..: Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nature Human Behaviour, 1(4):0064. (2017)