AIエージェントの挙動に関する共通認識の構築に向けた人とのコミュニケーションモデルの研究

2023年3月

福地庸介

別表５
⑶

主　論　文　要　旨

No.1

報告番号	㊙甲 乙 第　　号	氏　名	福地　庸介

主 論 文 題 名：
AI エージェントの挙動に関する共通認識の構築に向けた人とのコミュニケーションモデルの研究

（内容の要旨）

機械学習によって行動を学習する自律エージェント（AI エージェント）が、技術の発展にともなって複雑な挙動を扱えるようになっている。しかし、既存の AI エージェントの設計では、エージェントと共存するユーザや周囲の人とのコミュニケーションが十分に考慮されていない。AI エージェントの多くは事前に与えられた価値基準を最大化することに特化しており、それをもとに学習した行動方策は、実際の場面で人が持っている評価基準や期待とは必ずしも一致しない。真に人に利する存在となるために AI エージェントは、学習した方策に固定的に従うのではなく、ユーザが期待する挙動を適宜認識できるようにすることが重要である。また、AI エージェントが人の期待を認識するだけでなく、人も AI エージェントの挙動を正しく把握できるようにする必要がある。エージェントの挙動を人が把握できなければ、エージェントが予測不能となり、意図せぬ動作や思わぬ事故に繋がる。しかし、特に深層強化学習モデルは大量の数値パラメータで表現され、多くの場合で人が直接理解できる表現を持たない。こうした学習モデルブラックボックス化の問題は、モデルの複雑化や大規模化とともに深刻さを増している。

本論文の目標は、人と AI エージェントの共存に向けて、AI エージェントの挙動に関する相互理解を構築する挙動アライメントを達成することである。そして、コミュニケーションを通じて挙動アライメントを達成するための 2 つのモデルを提案する。2 つのモデルに共通するのは、AI エージェントと人との間で、目標や信念といった心的状態を帰属し合う〈心〉の読みあいの中で、両者の間に存在する情報の差異（非対称性）を明示的に組み込むことで、効果的なコミュニケーションを実現している点である。

第 1 に提案する「期待されるエージェント」モデルは、人が AI エージェントに達成を期待する目標を推測しながらコミュニケーションを行う過程をモデル化したものである。このモデルには、人の目標は AI エージェントの目標と必ずしも一致しないという目標の非対称性の存在が組み込まれている。目標の非対称性を考慮することで AI エージェントは、人から与えられる指示を正しく解釈することができる。さらに、解釈した指示の語彙を流用することで、AI エージェントがどのように動こうとしているかを人に伝達できるようになる。評価実験の結果、「期待されるエージェント」モデルが、指示の背後にある目標を正しく推定できること、推定した目標を元に解釈した指示の語彙を AI エージェントの動きの伝達に利用することで、人が AI エージェントの挙動を精度よく予測できるようになることが示された。

別表５
(3)

主　論　文　要　旨

No. 2

第２に提案する「推測されるエージェント」モデルは、AI エージェントの動きを見た人がエージェントに対して帰属する目標を、エージェントの側から推測しながらコミュニケーションを行う過程をモデル化している。この時、人と AI エージェントが異なる視界から環境を観測しているという観測の非対称性を考慮することで、人が AI エージェントに帰属する目標を正しく推測できる。　さらに、このモデルを応用することで、AI エージェントの目標を伝達する動き（表意動作）を生成できる。評価実験の結果、「推測されるエージェント」モデルが生成する表意動作によって、人が AI エージェントの目標をより早く、正しく推測できるようになることがわかった。

別表６
（３）

Keio University

Thesis Abstract

No.

Registration Number	☑"KOU" □ "OTSU" No. *Office use only	Name	Yosuke Fukuchi

Thesis Title

A Study on Communication Models for Behavior Alignment Between Humans and AI Agents

Thesis Summary

With the development of machine learning, AI agents, which autonomously learn actions through machine learning, are becoming capable of making complex decisions. However, their lack of ability to communicate with humans makes human-agent coexistence challenging. As a result of poor communication regarding the agents' behavior, they cannot behave as humans expect them to. The behavior of AI agents with machine learning depends on training datasets or values defined by designers, so it does not always match users' expectations. Through communication, AI agents should be able to understand what users expect of them to align their behavior with user expectations. In addition, AI agents should be able to communicate what they will do because their black-box decision-making modules prevent users from understanding their future behavior, which can lead to unintended behaviors and serious accidents.

This thesis presents two studies that aim at behavior alignment, i.e., aligning how AI agents behave and how humans expect them to behave through communication. Each study proposes a communication model to enable an AI agent to achieve behavior alignment. The two studies have in common that they exploit "mind"-reading phenomena between humans and AI agents and take into account an information asymmetry between them.

The first study proposes an expected-agent model, which enables an AI agent to interpret human instructions to it and convey what the agent will do. By taking into account an asymmetry between the goals of a human and an agent, the agent can correctly interpret hu- man instructions. Moreover, the expected-agent model learns the vocabulary used in human instructions and diverts it to convey what the agent will do. Experimental results show that explanations of an AI agent's future behavior generated by the expected-agent model can reduce errors in human predictions of the agent's future location.

The second study proposes an inferred-agent model, which infers an AI agent's mental state attributed by a human observing the agent's motion to enhance behavior alignment communication. The model effectively handles an observation asymmetry between a human and an agent to infer an attributed mental state and generates legible motion, a motion that conveys the agent's goal. Experimental results show that the inferred-agent model can enable human observers to quickly infer an agent's goal by taking into account an observation asymmetry.