氏　　名　中橋 亮

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2324 号

学位授与の日付　2022 年 3 月 24 日

学位授与の要件　複合科学研究科 情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Human-Agent Teaming with Implicit Guidance

論 文 審 査 委 員　主　査　山田 誠二
情報学専攻 教授
高須 淳宏
情報学専攻 教授
稲邑 哲也
情報学専攻 准教授
市瀬 龍太郎
情報学専攻 准教授
荒井 幸代
千葉大学 大学院工学研究院 教授

（様式３）

博士論文の要旨

氏　　名 中橋 亮

論文題目 Human-Agent Teaming with Implicit Guidance

Over the course of AI research history, the development of autonomous agents that can collaborate naturally with humans is one of the ultimate goals and has long been a significant issue. Among the many types of collaboration between humans and autonomous agents, we focus in this work on the collaborative problem in which humans and autonomous agents work together to achieve one task. We call this “Human-Agent Teaming”.

The most straightforward approach for Human-Agent-Teaming is to have the agents concentrate on supporting the humans.

In this approach, an agent infers a human's goals or intentions and takes the action that is most preferable for helping to achieve the goal.

However, as these agents cannot modify the human's plan, the ultimate success or failure of the collaborative task depends solely on the human's ability to plan. In other words, if the human sets the wrong plan, the performance will suffer.

Furthermore, humans generally have bounded rationality due to their cognitive and computational limitations, which makes it difficult for them to efficiently come up with optimal plans except for very easy tasks.

As such, the performance of agents in this scenario is limited.

The solution to this problem is to have the agents, who do not have cognitive or computational limitations, make optimal plans and then guide the humans to follow the plans.

In the most naive approach, the agent explicitly guides the humans to action.

However, if agents abuse such explicit guidance, humans may lose their autonomy, i.e., their sense of control regarding their own decision-making.

As a result, humans may think that the agent is controlling them.

Such an impression makes the Human-Agent Team less attractive.

The best way to circumvent this is to have agents guide humans while enabling them to retain their autonomy. To this end, we focus on “Implicit Guidance” offered through behavior.

The agent will expect the human to infer its intentions from its behavior and discard any plans that do not match what they infer the agent is planning.

Under this expectation, the agent acts in a way that makes it easy for the human to find the best (or at least better) plan for an optimal performance.

Implicit guidance of this nature should help humans maintain their autonomy, since the discarding of plans is a proactive action.

This dissertation consists of three studies examining our methodology for autonomous agents to use implicit guidance for Human-Agent Teaming.

The first study details the basic framework for implementing collaborative agents based on implicit guidance and demonstrates the advantage of utilizing the agents in this way.
This framework extends the existing planning approach by equipping agents with the ability to consider the Theory of Mind.
The Theory of Mind refers to the human cognitive function of inferring the goals or intentions of others on the basis of their behavior.
By utilizing this function, an agent can control human inference of the agent's intention and guide them to better plans.
We conducted an experiment in which participants were asked to perform a simple synthetic task by collaborating with several kinds of autonomous agents, including the agent with implicit guidance.
Our findings showed that the agent with implicit guidance could achieve a balance between successfully performing the task and maintaining human autonomy.
The second and third studies extend the framework to more realistic problems.
In the second study, we introduce the ``Plan Predictable Bias'' into the existing Theory of Mind modeling.
This is kind of the "bounding rationality" of human cognition and the bias that inferrers tend to infer others' intentions to make the inference easier for them.
We conducted an experiment in which participants were asked to infer the agents' intentions from their behavior in a complex synthetic task.
Our findings showed that the Theory of Mind model with Plan Predictable Bias matches human cognition better than the existing Theory of Mind.
The third study extends the planning algorithm to the more realistic situation that the human has specific information about the reward that is unknown to the agent.
In this case, the agent cannot initially make the best plan, so it has to infer the specific information from the behavior of the human.
We implemented our implicit guidance concept in the existing collaborative planning algorithm, which expects the humans to show their intention and [the agents to?] infer it.
We conducted an experiment in which participants were asked to achieve a complex task by collaborating with several kinds of autonomous agents, including our extended agent.
Our findings showed that our framework with the extended agent improved the performance in achieving the collaborative task.
Despite several limitations, this dissertation contributes to fostering a more prosperous and natural relationship between humans and artificial intelligence.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　中橋 亮

Title
論文題目　Human-Agent Teaming with Implicit Guidance

本学位論文は、“Human-Agent Teaming with Implicit Guidance”と題し、全 8 章から構成され，英語で書かれている．

第 1 章 Introduction では，本研究の目的が，人間-AI 協調系において人間に行動を観測してもらい，その意図を推定させることで全体として効率的かつ人間にとって快適な協調行動を実現できる AI エージェントの開発であることが説明されている．そして，研究全体の概観，本論文の構成が述べられている．

第 2 章 Related Work では，マルチエージェントプランニング／協調プランニング，人間の認知モデルを利用したプランニング，ガイダンスを用いた協調プランニング，そして協調プランニングのための人間の印象評価について，関連研究が紹介されている．

第 3 章 Background では，本研究の理論基盤となるいくつかの概念，枠組みについて説明されている．具体的には，マルコフ決定過程 MDP，部分観測マルコフ決定過程 POMDP を始めとする様々なマルコフ決定過程が説明されている．そして，MDP および POMDP を利用したプランニングアルゴリズム，協調逆強化学習 CIRL，CIRL のためのプランニング，心の理論のベイジアンモデリング，人間の合理性モデル，そしてベイジアン心の理論が説明されている．

第 4 章 Human-Agent Teaming では，人間とエージェントが協力して一つの問題を解くために協調する系である人間-エージェントチーミングについて，その基本的概念が説明された．続いて，人間-エージェントチーミングの定式化について述べられ，従来の自律協調エージェントの多くは人の行動から人の意図や目的を推測し，推測した目的をサポートする supportive agent であることについて説明された．

第 5 章 Planning with implicit guidance では，行動によって人間に意図を伝える非明示的ガイダンスと通信により意図を伝える明示的ガイダンスの違いが説明された．その後に，非明示ガイダンスであるエージェントの行動系列のプランニングを含んだ全体的プランニングアルゴリズムが，ベイジアンモデリングをベースに開発されている．続いて，テストベッドである協調タスクの説明があり，そのテストベッドでの評価実験，その結果と分析が行われている．

第 6 章 Extension of planning with implicit guidance for complex task では，心の理論のベイジアンモデリングをより複雑な環境へ適用するために，タスクの具体例，問題のノーテイション，完全に精緻な推定を行う完全逆プランニング，そして提案手法である人間の認知的限界である限定合理性を導入したプラン予測性モデルが説明された．続いて，実験の設定，実験の実施方法，そして結果の分析について述べられた．また，プランニングアルゴリズムの拡張として，予測性バイアスを適用したプランニングアルゴリズム開発とその評

価実験，実験結果についての議論が説明されている．

第 7 章 General Discussion では，研究全体における有効性，限界，非明示ガイダンスエージェントが有効な環境に関する議論，そして潜在的な応用分野が説明された．

第 8 章 Conclusion では，本研究の全体的な構想，目的，方法そして評価について総括的にまとめられている．

公開発表会では，博士論文の章立てに従って発表が行われ，その後に行われた論文審査会及び口述試験では，審査員からの質疑に対して適切に回答がなされた．

質疑応答後に審査委員会を開催し，審査委員で議論を行った．審査委員会では，出願者の博士研究が，人間に行動を見せることで AI エージェントの目的推定を促すことを含めた行動プランニングをベイジアンモデルに基づいて定式化し，その計算方法も提案していることが評価された．

以上を要するに本学位論文は，今後重要になる人間に受け入れられる AI 技術を人間の快適さを担保した上で実現する非明示ガイダンスを利用した協調エージェントの方法論を開発して，綿密な評価を行った研究であり，これからの社会への AI 応用・導入の可能性を大きく広げる先駆的研究である．また，本学位論文の成果は，学術雑誌論文 1 件，査読付き国際会議論文 1 件として発表され，学術的な貢献も認められる．

以上の理由により，審査委員会は，本学位論文が学位の授与に値すると判断した．