氏　　名　岡村 和男

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2191 号

学位授与の日付　2020 年 9 月 28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　Adaptive Trust Calibration in Human-AI Cooperation

論 文 審 査 委 員　主　査　教授　山田　誠二

准教授　市瀬　龍太郎

准教授　相原　健郎

准教授　稲邑　哲也

准教授　前東　晃礼
静岡大学　学術院融合・グローバル領域

（様式３）

博士論文の要旨

氏　　名　岡村 和男

論文題目　Adaptive Trust Calibration in Human-AI Cooperation

Recent advances in AI technologies are dramatically changing the world and impacting our daily life. The application areas are rapidly expanding, such as autonomous cars, industrial robots, medical services, and various web services. Human users essentially need to cooperate with AI systems to complete tasks as such technologies are never perfect.

One key aspect of human-AI cooperation is that human users should trust AI systems, just as humans normally do with other human partners. The presence and absence of trust definitely impact human behavior and the outcome of cooperation. For optimal performance and safety of human-AI cooperation, the human users must appropriately adjust their level of trust to the actual reliability of AI systems. This process is called ``trust calibration''. Users often fail to calibrate their trust properly and end up in a status called ``over-trust'' or ``under-trust'' in dynamically changing environments in which an AI's reliability may fluctuate. Poorly calibrated trust can be a major cause of serious issues with safety and efficiency.

A large number of existing studies on trust calibration emphasize the importance of system transparency to maintain appropriate trust. They claim that appropriate trust could be developed if an AI system provides enough information for a human user to obtain a good understanding of the system. Their primary goal is to avoid over-trust or under-trust, not to deal with improper trust calibration.

Trust is notoriously hard to measure as it is a psychological construct. Self-reported scales of trust that are widely used in most trust literature are too intrusive to use during task executions. Extensive studies have been conducted to examine the factors influencing trust. Although the findings of these pieces of literature revealed the diversified latent structures of human trust, they suggest that it would be difficult to influence human trust intentionally just by manipulating these factors. Thus, both measuring and influencing trust are challenging issues.

This dissertation focuses on the problem of over-trust and under-trust in human-AI cooperation by exploring two research questions: (1) Can we detect if a user is over-trusting or under-trusting an AI system? (2) Can we mitigate a user's over-trust or under-trust?

We approach the research challenges with a behavior-based trust measurement to capture the status of calibration. Human-AI cooperation is defined as a series of actions taken by a human user and an AI system working on repeated selection problems to decide on either AI execution or manual execution for better performance. A method of adaptive trust calibration is proposed, including a formal framework for detecting improper trust calibration; cognitive cues called ``trust calibration cues''; and a technical architecture of human-AI cooperation with a concept called trust calibration AI.

Three empirical studies were done to evaluate the proposed method. We designed two experimental tasks for human-AI cooperation: a pothole inspection task and a continuous cooperative navigation task. Three online experiments using a simulated drone environment were conducted. We observed both the status of over-trust and the under-trust for the participants of all three experiments. The results of the first empirical study demonstrate that our proposed method has significant effects on changing human behavior in the case of over-trust. A verbal cue showed the largest effect amongst the other cues of visual, audio, and anthropomorphic. The second empirical study shows that the proposed method also works well under dynamic trust changes of ABA and BAB, where A and B mean over-trust and under-trust. The third empirical study indicates that the proposed method is effective in a continuous real-time task involving navigating a semi-autonomous drone. This result can open the possibility of applying the proposed method to practical real-time applications such as autonomous driving. We also discuss a possible extension to the framework with expected utility functions to incorporate trust factors other than performance.

The results of the empirical evaluations indicate that the proposed method could detect and mitigate the status of improper trust calibration; therefore, we conclude that our proposed method provides a reasonable basis for answering the two research questions.

As the proposed method is based on a simple and task-independent framework, it could be applied to many application situations. Despite several limitations, this dissertation contributes to providing a basic framework for managing trust calibration, leading to better interaction designs for human-AI cooperation.

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　　岡村 和男

Title
論文題目　　Adaptive Trust Calibration in Human-AI Cooperation

本学位論文は，“Adaptive Trust Calibration in Human-AI Cooperation” と題し，全 5 章から構成され，英語で書かれている．第 1 章 “Introduction”では，本研究の背景となる人間と AI の協調系について，自動運転，AI による意思決定支援などの例を基に説明している．加えて，人間-AI 協調タスクにおける適切な信頼の構築のために，過信・不信を是正する信頼較正の重要性に触れている．また，信頼較正の解決方法について説明され，本研究の目的が「人間と AI の適切な信頼関係を構築するために，人間の信頼較正を促進する信頼較正 AI の提案と実験的評価である」ことが説明されている．そして，論文の全体構成について述べられている．

第 2 章“Related Work”では，本研究で扱う信頼の概念の体系化，位置づけ，信頼に影響する要因の整理,そして信頼のモデル化と計測方法について関連研究を説明している.また，信頼工学，ヒューマンロボットインタラクションと信頼較正の関係も考察されている．

第 3 章“Adaptive Trust Calibration”では，本研究の主たる貢献として，AI が人間の過信・不信状態を検出し，較正キューの表出により適応的に信頼較正を促す枠組みである，適応的信頼較正アルゴリズムについて説明している．タスクの成功確率による信頼の定式化と信頼方程式を基に合理的に振る舞う人間の人間-AI 選択行動から過信・不信を判定するアルゴリズムについて述べている．

第 4 章“Empirical Studies”では，提案手法の評価として，ドローンシミュレーションを使った実験と実験結果の分析について説明している．まず，ドローンによる道路検査タスクの人間-AI 協調タスクにおいて実施した 2 種類の参加者実験とその結果の分析について説明している．これら実験では，人間に較正を促す刺激である較正キューについて様々な候補が比較検討され，動的に変化する過信・不信状態へ提案方法が対応可能であることを確認している．次に連続的に変化するタスクへ提案方法が応用できることを示すために行われた，ドローンの人間-AI 協調ナビゲーションにおける参加者実験の計画，実施，そして実験結果の分析について説明している．

第 5 章 “Conclusion”では，全体的な議論・考察がされている．また，信頼の表現の拡張について検討し，その応用例について考察している．最後に，論文全体の結論と今後の発展性について説明さている．

以上を要するに本学位論文は，人間-AI の協調タスクにおいて，人間の AI に対する過信・不信を較正する枠組みを提案し，様々な実験によって，その提案方法を評価したものである．

公開論文発表会では，出願者は約 45 分で博士論文の内容を説明し，続いて 20 分程度の質疑応答と審査員のみでの口述試験が行われた．以上における審査員からの質疑に対して

出願者は適切に回答した．

その後，審査委員会が開催され，審査委員で議論を行った．審査の結果，出願者は情報学分野の十分な知識と研究能力を持つと認められ，また研究内容は学位論文として十分なレベルの新規性，有効性があると認められた．また，本学位論文の成果は，学術雑誌論文1編，査読付き国際会議論文3編（すべて筆頭著者）として発表されている．以上の理由により，審査委員会は全員一致で，本学位論文が学位の授与に値すると判断した．