信頼される AI の基盤技術
2020 年度採択研究者

2021 年度
年次報告書

西田 知史

情報通信研究機構 未来 ICT 研究所脳情報通信融合研究センター／大阪大学 大学院生命機能研究科
主任研究員／招へい准教授

脳情報に基づいた AI の信頼性評価技術の開発

§1. 研究成果の概要

本研究計画では、信頼されるAIの基盤技術開発に資することを目的として、認知神経科学の方法論と知見を活用した2つのアプローチでAIの信頼性に関する研究を進めている。1つ目のアプローチは、人間がAIを信頼するとは何かという根本的な問いに対する認識論的説明を得ることを目的とした、AIへの信頼性を生み出す脳内基盤の探究である。2つ目のアプローチは、信頼されるAIに不可欠な要因として人間らしい判断を行うAIを実現することを目的とした、脳情報のモデルをAIへ取り入れる技術の開発と検証である。1つ目のアプローチに対する本年度の研究では、AIに抱く負のイメージがAI生成画像に対する人間の魅力度判断にもたらす影響を確認した。また、その際に機能的磁気共鳴画像法(fMRI)によって計測した脳活動を分析し、負のイメージが視覚野を含む広い脳領域の情報に変容をもたらすことを示した。2つ目のアプローチに対する本年度の研究では、個人脳の情報処理を定量化したモデルをAIに融合したうえで、そのAIが下す判断と脳活動から読み取った人間の判断の関係性について分析を行った。その分析の結果、個々人の脳情報処理モデルを融合したAIの判断は、脳活動から読み取った判断の個人差を反映することが明らかになり、我々の開発したAIは脳情報の個人差をうまく模倣することを確認した。以上の成果は、AIに対する信頼性の障壁となる負のイメージが人間の認知プロセスへ影響を与える脳内機序に関する理解を得るとともに、AIを人間らしい個性を持ったエージェントへと進化させるための基盤となりうる技術を開発したことを示しており、信頼されるAIの研究開発に多大な貢献をもたらすことが期待できる。