脳情報に基づいたAIの信頼性評価技術の開発

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　主任研究員）

研究領域「信頼されるAIシステムを支える基盤技術」
（研究総括：有村　博紀、2020年度発足）

研究の概要

信頼できるAIの開発にあたり、AIが扱う情報の安全性といった客観的信頼性とともに、AIの提示情報に対して人間が感じる主観的信頼性の向上も重要な課題となる。本研究では、人間がAIの提示情報に主観的信頼性を感じる脳内メカニズムを脳計測実験を用いて解明し、それに基づいて脳活動から主観的信頼性を解読してAIを評価する革新的技術を開発する。また、シミュレートした脳活動を解読する手法を融合し、大幅な脳計測コスト削減を実現する。

提案研究終了時の達成目標

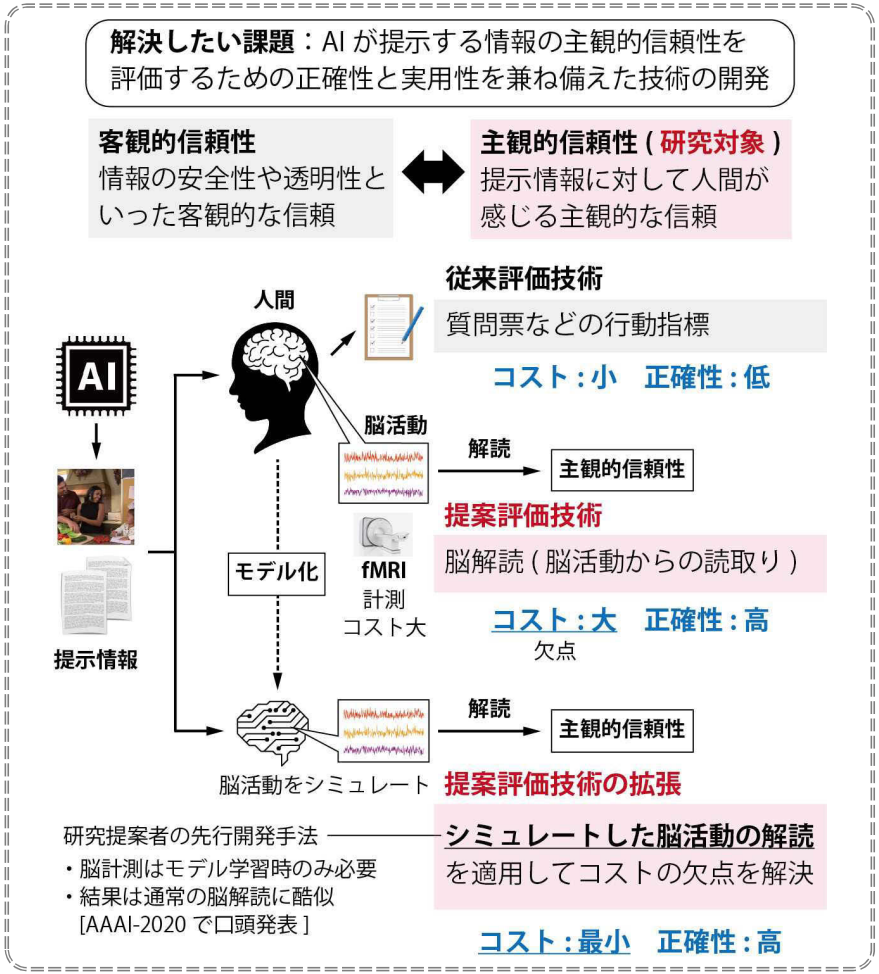
信頼性評価を実データで成功させ、従来の行動指標に対する優位性を示す。

提案研究の独創性、新規性・優位性

従来研究はAI（ロボット）の主観的信頼性を質問紙等の行動指標で評価している [Hancock+11]。脳活動の解読（脳解読）は、より正確な評価を与えうるが [Falk+12]、従来の脳解読は知覚内容 [Haxby+14]や好悪 [Nishida+20] などに限られ、主観的信頼性の脳解読は世界初となる。さらに、シミュレートした脳活動の解読手法によりコストを大幅削減すれば、高い正確性と実用性を持つ、信頼できるAIの開発にブレイクスルーをもたらす画期的技術が生まれる。

提案研究の挑戦性

AIの主観的信頼性を生み出す脳内メカニズムの解明と、それを利用した脳解読は世界初の挑戦的な試みであり、その成果は脳科学分野とAI分野に多大なインパクトを与える。



研究の将来展望

（1）学術研究としての、さきがけ研究成果の将来展開

本成果は、AIの主観的信頼性の標準的評価技術となり、信頼できるAIの研究開発に底上げをもたらす。また、人間が情報を信頼する脳内機序の理解にもつながり、フェイクニュースやサイバー詐欺に騙される認知的要因の探究など、豊かで安心な情報社会実現のための重要な研究に発展しうる。

（2）さきがけ研究成果と社会との将来の接点（新技術の創出・知的財産権の取得及び活用、又は社会普及・社会受容等）

本研究の開発技術は高い新規性を持つため知財化が期待できるとともに、信頼できるAIのデザインにおける新たな国際標準規格を提供しうる。