

## 研究課題別事後評価結果

1. 研究課題名： 脳情報に基づいた AI の信頼性評価技術の開発

2. 個人研究者名

西田 知史（情報通信研究機構 脳情報通信融合研究センター 主任研究員）

3. 事後評価結果

本研究は、AI の提示情報に対する人間の主観的信頼性のメカニズムを解明し、AI に対する主観的信頼性を脳活動から解読・評価するための技術基盤を開発する、極めて挑戦的かつ独創的な取組である。このために、人間の主観的信頼性を生み出す脳内メカニズムに着目し、最先端の脳科学の知識と脳活動計測・解読技術に基づき、以下の2つのアプローチで研究を進めた。

1つ目のアプローチとして、人間による AI への信頼メカニズムの解明に関し、認知神経科学の方法論と知見を用いて、AI への信頼性を生み出す脳内基盤の探究を行った。AI 生成画像に対する人間がもつ魅力度判断における脳内機序を、最新の機能的脳活動計測機器である fMRI を用いて、約 60 名の被験者による大規模計測実験を行うことで、AI 生成情報に対する人間の負の価値判断バイアスと脳応答の関係を明らかにした。さらに同規模の実験で、AI 不安感の個人差を生み出す脳内基盤を分析した。2つ目のアプローチとして、人間らしい判断を行う AI の実現に関して、脳情報のモデルを AI へ融合する技術である脳融合深層ニューラルネットワーク（脳融合 DNN）の研究・開発を行った。映像の意味認知タスクの実験では、脳融合 DNN が個人差を反映した高精度な判断能力をもつことを示唆する結果を得た。

これらの成果は、脳神経科学、計算生物学、人工知能分野のトップクラスの国際会議・学術雑誌の論文として発表され、1 件の国際学会賞と 2 件の国内学会賞を受賞するなど、高い評価を受けた。これらの成果は、最先端の脳神経科学と AI 技術の融合により、AI の提示情報に対する人間の主観的な信頼性の仕組みを解明し、人間の個性を反映した AI システムの実現の可能性を示した点で画期的な研究であり、その学術的意義は極めて大きい。また、さきがけ研究領域においては、脳・人間科学と情報科学の境界を越える異分野間の議論と連携を積極的に進め、将来の信頼される AI に関する議論をリードし、深めてくれた。本研究は、未来の AI 社会において人と共存・協働する信頼される AI 技術の実現に向けた重要な一歩となると思われる。脳情報と AI 技術の融合の基礎研究を確実に進めるとともに、将来的な実用化に向けた継続的な取組も必要となるであろう。信頼される AI の地平を切り開く、脳 AI 融合の第一人者として、今後のさらなる飛躍を期待する。