信頼される AI の基盤技術　　2022 年度
2020 年度採択研究代表者　　年次報告書

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

# 脳情報に基づいた AI の信頼性評価技術の開発

## 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 に抱く不安感の個人差を生み出す脳内基盤に関する知見を得た。被験者 59 名に 6 つの項目から成るアンケートに回答してもらい個々人の AI に抱く不安感を評価した。一方、同被験者が MRI 装置内で AI・ロボットが登場する多様な映像を視聴している際の脳活動を計測した。そして、アンケート回答の被験者間非類似度を評価するとともに、脳活動の非類似度もボクセル（fMRI の最小計測単位）ごとに評価し、これら 2 つの非類似度におけるピアソン相関を分析した。その結果、いくつかの脳領域における活動の非類似度が AI に抱く不安感の非類似度と相関することが分かり、特に上側頭皮質における相関が顕著であった。上側頭皮質は生物・非生物の判別や他者の感情推定などに関与することが知られており、上側頭皮質において AI に抱く不安感の個人差の反映が見られたことは妥当な結果だといえる。また、2 つ目のアプローチに対する本年度の研究では、これまでの研究では視聴覚を扱う深層ニューラルネット（DNN）への脳情報融合を行ってきたが、新たに言語 DNN への脳情報融合を行い、モダリティの拡張を試みた。結果として、言語 DNN の一種である BERT に脳情報を融合したところ、パターン認識課題における認識性能の向上、認識結果における脳への近接、脳情報の個人差の反映という視聴覚 DNN への脳情報融合でも見られていた 3 つの効果が言語 DNN でも見られることが確認できた。つまり、言語 DNN においても脳情報融合が効果的に機能することが示された。以上の成果は、AI に対する信頼性の障壁となる不安感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。

【代表的な原著論文情報】


1) Nishida S. Behavioral and neural evidence for the underestimated attractiveness of faces synthesized using an artificial neural network. bioRxiv:2023.02.07.527403, 2023.
2) Matsumoto Y, Nishida S, Hayashi R, Son S, Murakami A, Yoshikawa N, Ito H, Oishi N, Masuda N, Murai T, Friston K, Nishimoto S, Takahashi H. Disorganization of semantic brain networks in schizophrenia revealed by fMRI. Schizophrenia Bulletin, 49(2):498–506, 2023.
3) Kawasaki H, Nishida S, Kobayashi I. Hierarchical Processing of Visual and Language Information in the Brain. Proceedings of AACL-IJCNLP 2022, 405–410, 2022.
4) Kawahata K, Wang J, Blanc A, Maeda N, Nishimoto S, Nishida S. Decoding Individual Differences in Mental Information from Human Brain Response Predicted by Convolutional Neural Networks. bioRxiv:2022.05.16.492029, 2022.