Vol.2021-NL-247 No.5
2021/3/3

AIの安全・信頼と自然言語処理を考える

荒井　ひろみ[1,a)]

概要：AI の安全・信頼のために，プライバシー保護や公平性・説明性が重要視されている．パーソナルデータなどの機微な情報を利用するために，プライバシーリスクの評価や，データを安全に分析・収集・共有するためのプライバシー保護技術がある．また近年機械学習モデルの公平性が着目されている．データに含まれるバイアスの影響により機械学習モデルが差別的な振る舞いをすることが問題視されており，バイアスの分析や調整を行う方法が提案されている．さらに複雑化する AI についてユーザーに説明をするために，機械学習モデルの振る舞いを説明する方法や，プライバシーポリシーの正確さやユーザビリティの検証などが行われている．本講演ではこれらの研究について講演者のこれまでの研究を交えつつ紹介し，このような技術を言語に適用する際の課題や，自然言語処理との連携の可能性を議論する．

[1] RIKEN Center for Advanced Intelligence Project
[a)] hiromi.arai@riken.jp