FY2023/2023年度

AI Safety and Reliability Unit

Hiromi Arai
人工知能安全性・信頼性ユニット　荒井 ひろみ

RIKEN Center for Advanced Intelligence Project

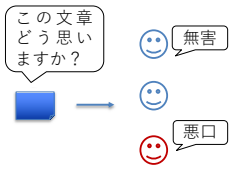
研究目標

AI利用における安全・信頼のための課題の検討および対処方法の開発

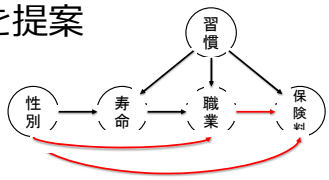
AIの公平性

AIが公平であるための、不公平な判断の評価や対処の方法を研究しています。特にデータの偏りがある場合の学習や、学習データにおけるアノテーションバイアスの分析、多様な判断がある場合の意見集約などを研究対象としています。

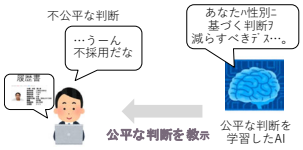
- ヒューマンコンピュテーションにおけるラベルのバイアス
- ヘイトスピーチ検出に向けたアノテーションの試案[荒井+20]
- 顔認証における公平性評価[大木+21]



- 因果関係を考慮した公平性[波多野+23]
  公平な因果関係と不公平な因果関係が存在する状況下において、AIの判断を公平化させる手法を提案



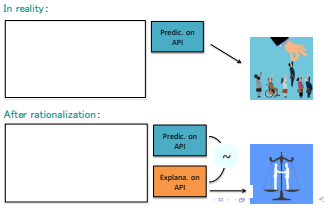
- より公平な判断をクラウドワーカーにさせるための機械教示[楊+21,Yang+24]
  各自の判断基準の偏りを具体的に示し公平な判断基準を人間に提示するAIを実装し、人の判断への影響を調査



説明可能AIの課題

AIの振る舞いを説明するにあたり、その課題や適切な方法についての研究をしています。

- 機械学習モデルの説明の偽装のリスク[Aïvodji+19, Aïvodji+21]
- 説明可能AIにまつわる議論の整理



情報環境の安全性とHCI

生成AIの登場などを受け、偽誤情報への対策の重要性は増しています。ファクトチェック行動の支援に関する研究をしています。

- ファクトチェック情報に対するクリック行動の分析[Tanaka+23]
- 情報提示方法によるクリック行動への介入[Tanaka+23]