FY2022/2022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

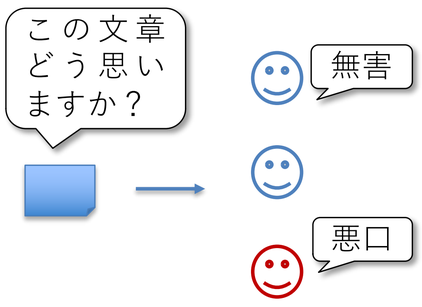
研究目標

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や、多様な判断がある場合の意見集約、有害な情報の判断などを研究対象としています。

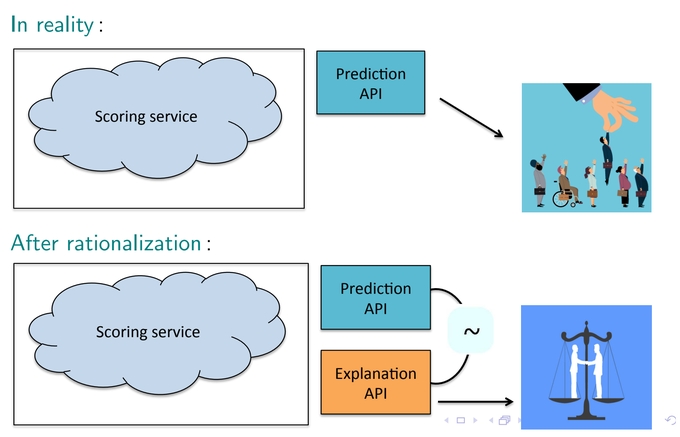
- ヒューマンコンピュテーションにおけるラベルのバイアス
- ワーカーにより公平な判断をさせるための機械教示[楊+21]
- ヘイトスピーチ検出に向けたアノテーションの試案[荒井+20]
- 顔認証における公平性評価[大木+21]



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
  - ユーザーのニーズが多様な領域においては対話型の対比的説明は有効と考えられるが，多重決定などが課題と考察[濱本+21]
  - 非専門家への目的帰属型説明の可能性についての議論[葛谷+22]



データプライバシー

- パーソナルデータの利用において、プライバシーを適切に保護するための方法として、プライバシー保護技術や健全な同意取得などの研究をしています。とくにプライバシーポリシーなどのデータの扱いのユーザーへの説明などを扱っています。
  - Contextual Integrityを用いたプライバシーポリシーの分析[荒井+20]
  - プライバシー保護異常検知[Arai+21]