FY2025/2025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

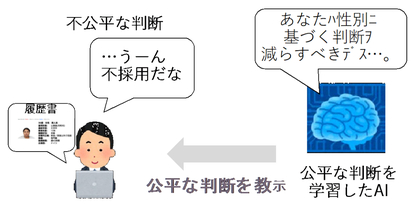
研究目標

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。特に偏りのあるデータからの学習や、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。

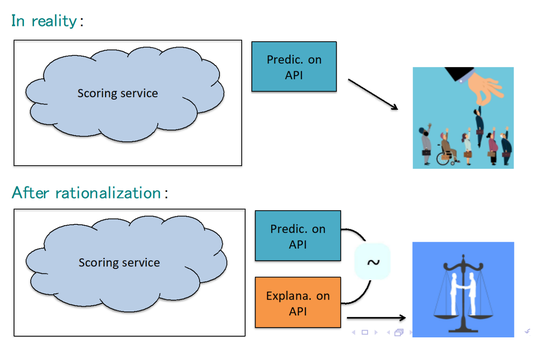
- ヒューマンコンピュテーションにおけるラベルのバイアス
- ヘイトスピーチ検出に向けたアノテーションの試案[荒井+20]
- より公平な判断をクラウドワーカーにさせるための機械教示[楊+21,Yang+24]
- 顔認証における公平性評価[大木+21]



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。

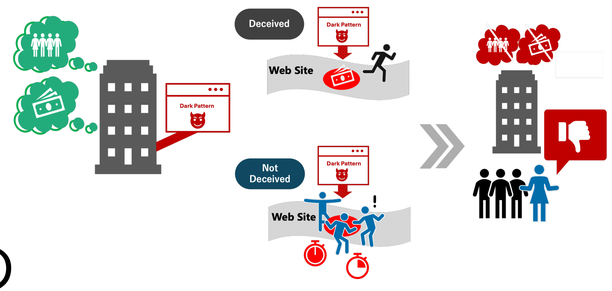
- 機械学習モデルの説明の偽装のリスク[Aïvodji+19, Aïvodji+21]
- 説明可能AIにまつわる議論の整理



Usable Security

ダークパターンやプライバシーポリシーのユーザビリティやリスクについて研究をしています。

- ダークパターン回避にかかるコストが企業への信頼を損なうリスク要因となり得る可能性[Kojima+24]
- Contextual Integrityを用いたプライバシーポリシーの分析[荒井+20]



情報環境の安全性

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]
- 情報提示によるクリック行動への介入[Tanaka+23][Tanaka+25]

また生成AIによる情報環境への影響についても研究をしています。

- 生成データの学習データへの混入の悪影響[Hataya+24]