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

背景：AIの予想外の振る舞いやAIが悪用されるリスクへの対応が課題
目的：AI利用におけるリスクの検証・対処方法の開発
　　プライバシー保護，AIの説明可能性，公平性，データのバイアスや有害性

AIの説明におけるミスリードの可能性

機械学習モデルの説明におけるfairwashingを検討
- Fairwashing-モデルがある倫理的な値(公平性)を実際は満たしていないが，満たしていると見せかけること

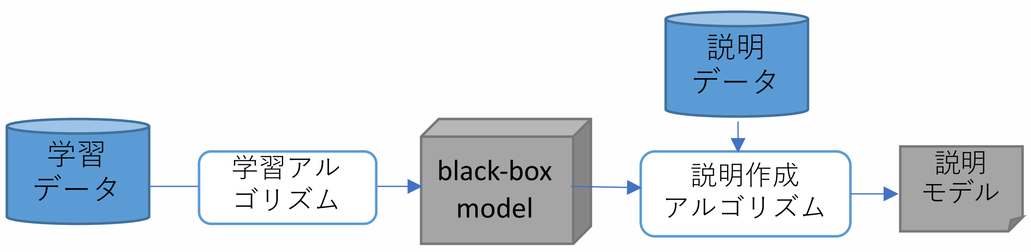


Fig.1 MLモデルのFairwashingの概要

Fairwasingの検出可能性を実験的に検討→検出は困難
- Fairwashingによって生じるFidelity loss
- Fairwashingの説明データへの依存性
- Fairwashingが標的とするモデルへの依存性

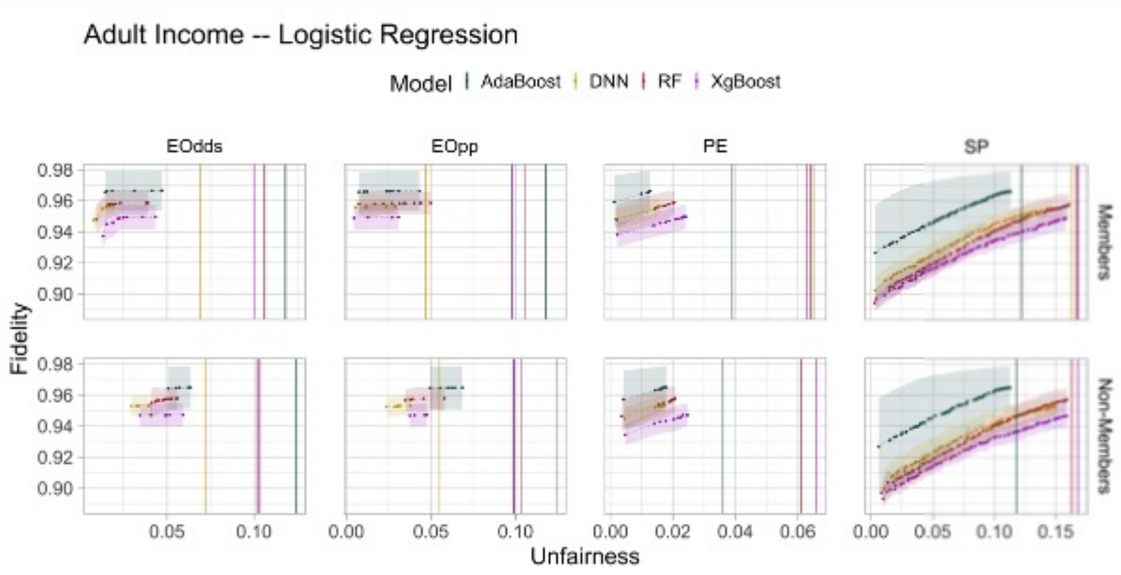


Fig.2 Fidelity-unfairness trade-offs

- 説明手法を操作することによりblack-boxモデルの動作についてユーザーをミスリードすることができる
- Fairwashingの検出は困難→作成プロセスに透明性をもたせる必要
- Fairwashingリスクの評価に可能な事後説明のunfairnessの取りうる範囲が使える可能性

(Aïvodji et. al.,, Neurips2021)

ヘイトスピーチの収集・分析

日本語の排外主義的なヘイトスピーチ候補ツイートに対し精緻化したガイドラインによるアノテーションを実施，コロナ禍でのヘイトスピーチの傾向やアノテーションの揺れについて考察した.

(荒井 et. al, 2021，和泉et.al, 2021)

説明可能AIにまつわる議論の整理

Mittelstadtらが提案する説明可能AIにおける対話型の対比的説明について検討，ユーザーのニーズが多様でありうる領域においては有効と考えられるが，多重決定などが課題と考察した.

(濵本 et.al.，JSAI2021)

顔認証における公平性評価

顔認証において学習データにおけるセンシティブ属性の偏りが顔照合及び識別精度に与える影響評価方法を検討

表現学習を用いた顔認証における公平性評価を検証
MORPH Datasetから人種/性別をセンシティブ属性とし，属性別の2サブグループの人口比率を変えて学習/テストデータをサンプリング，FMR，FPIRなどを評価した.

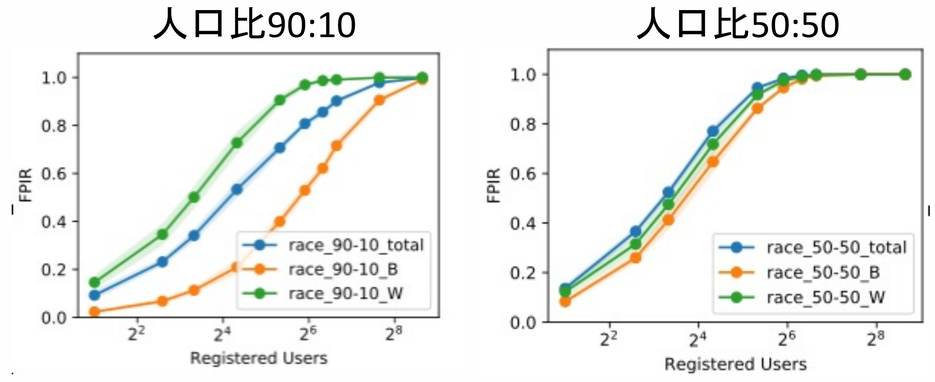


Fig. 3　属性別FPIRの評価．学習データの偏りが大きいほどFPIRの違いが大

- 人口が少ない属性集団は顔認証精度が低い傾向
- 適応的しきい値設定により少数属性グループの精度の改善可能性

(大木，荒井，CSS2021)

公平な判断の機械教示

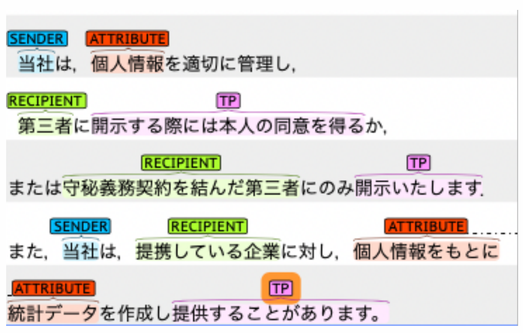
より公平な判断をクラウドワーカーにさせるために機械教示を適用，効果を検証している
ある評価者が行った評価結果に対し，公平性配慮型機械学習を用い公平なモデルを作成し，機械教示を行う.
疑似人物情報をもとにした年収予測タスクにおいて，機械教示の評価を試みた

(楊 et.al.，JSAI2021)

パーソナルデータ利用における健全な同意

プライバシーポリシーの課題
- 曖昧な表現や不正確な記述の使用
- 複雑で難解な長文のため読まれない

健全な同意に向けて
- プライバシーポリシー等利用者への説明の実態調査
- 説明の評価，情報提示方法の検討



文脈的完全性に基づいた情報抽出
[荒井+2020]などの方法を利用