大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

AIの公平性に関する企業リスクについて

•小林，佐藤，西山，藤村，八幡
•チームD3 (石川先生)

開発における問題点

昨今，AIは広く使われるようになり，AIが人を評価する際に公平な取り扱いが求められるようになった．しかし公平性と言っても様々な考え方・指標があり，①異なる指標がトレードオフになり公平性実現が困難に見えたり，②近年登場した公平性ツールキットは価値判断をユーザに委ねており，公平性実現に向けた明確な指針を示していないといった点が，開発における問題点として挙げられる．

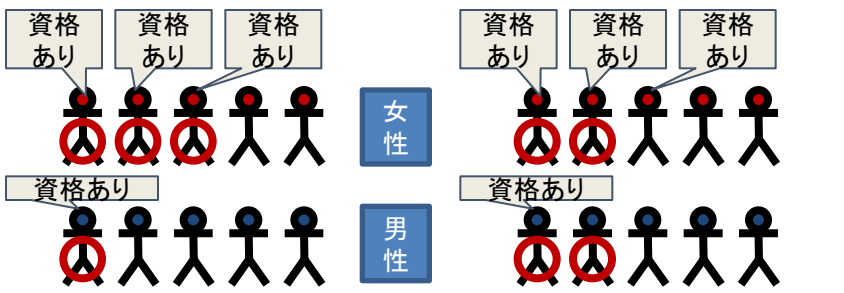
手法・ツールの適用による解決

そこで，①に関してはツールキットを使用して実験を行うことにより，一見トレードオフに見える異なる指標が同時に改善できるケースがあることを示した．すなわち公平性改善は可能であり，その努力を怠ると企業は法的リスクを負いかねないということだ．他方，公平性は技術者にとってまだまだとっつきにくい課題である．そこで②主に技術者を対象として公平性実現への取組みを支援するため，ツールキットのビギナーズ・ガイドを作成した．

問題点

手続きの公平性
(指標: EO=Equalized Odds)
男女問わず資格のある人は合格

結果の公平性
(指標: DP=Demographic Parity)
資格関係なく男女同割合合格



ツール名	ツール思想
Fairlearn (マイクロソフト)	ユーザが公平性に関連する害を評価し，様々な緩和戦略の影響を検討し，そのシナリオに適したトレードオフを行えるようにすることが目標.
AIF360 (IBM)	様々な公平性指標と緩和アルゴリズムが存在する中，その中から適切なものを選ぶための議論の出発点として利用することを想定.
Google	公平性指標として何を採用するかは簡単な問題ではなく，他のツールキットと同様，

公平性実現上の問題点

①手続きの公平性と結果の公平性を両立するのは難しい．公平性指標もトレードオフの関係に思える．よって企業も真面目に取り組みづらい

ツールキットの問題点

②ツールキットはそれぞれガイドラインを持っているが，最終的は判断はユーザに委ねており，具体的な指針がない

①公平性改善事例

緩和アルゴリズム	accuracy	DP	EO
Base	0.844	0.15	0.08
Exponentiated Gradient (DP)	0.821	0.02 ✓	0.29
Exponentiated Gradient (EO)	**0.839**	**0.11** ✓	**0.02** ✓
Grid Search(DP)	0.833	0.07 ✓	0.17
Grid Search(EO)	0.844	0.15	0.08
Threshold Optimizer(DP)	0.825	0 ✓	0.33
Threshold Optimizer(EO)	**0.829**	**0.09** ✓	**0** ✓
CorrelationRemover	0.841	0.19	0.32

✓: 指標が改善

- **DP (結果の公平性) と EO (手続きの公平性) の双方がベース手法に比べ改善する手法が2つ存在している**
- **Accuracy（精度）の低下も小さい**

②ビギナーズ・ガイド作成

CATION PROGRAM FOR TOP SOFTWARE ENGINEERS

目次

Point.1
マイクロソフト/IBM/Google 3社のツールキットを俯瞰して，共通する要素を抽出

Point.2
数ある公平性指標の中でも，DP/EOの2つを明示的に推奨

Point.3
手続きの公平性と結果の公平性が共に改善する場合があることを明示

期待効果

プロダクトオーナーやエンジニアが

- **AIの公平性に視野を向ける**
- **公平性ツール使用時の効果を確認し，ツール未使用時のリスクを認識する**
- **フェーズに合わせ，適切なツールを使える**
- **公平性を追求しても精度がさほど落ちないことがわかる**