Fujitsu AI Ethics for Fairness

富士通株式会社
富士通研究所 AIトラスト研究センター

背景

AIによって新たに生じる問題

FUJITSU

行政

顔認識の人種による高い誤検知率から
米国各州で使用禁止

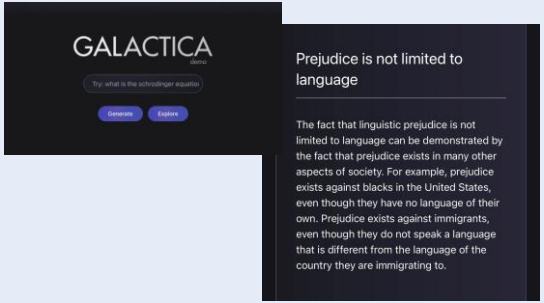
https://www.ajl.org/
https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/

人材採用

人材採用AIが女性の評価を
不当に下げていると発覚

https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN

「生成AI」



ネット公開した対話AIが、不正確さや
差別的内容で、3日間で非公開に

https://arstechnica.com/information-technology/2022/11/after-controversy-meta-pulls-demo-of-ai-model-that-writes-scientific-papers/
https://twitter.com/thai101/status/1592752955694153728

信用の失墜による事業の中止や従業員エンゲージメントの低下のリスク

AIを安心して提供できる仕組みが必要 → 「AI倫理」

AI倫理を取り巻く環境

● 欧州では、AI規制の法制化の準備が進んでいる

Nov. 2018
Ethical Framework for Good AI Society
4 PEOPLE

Apr. 2019
Ethics Guidelines for Trustworthy AI
「信頼できるAIのための倫理ガイドライン」

Apr. 2021
AI Act
（AI法規制案）
EUROPEAN COMMISSION

2025~
AI規制施行

欧州法制化案　AI Act	
時期	2025以降
罰金	€30Mまたは売上の6%
対象	提供者・販売者・使用者

● EUの法規制案によって、今後世界各国市場にも影響を与える

参考　EUルールが日本の法規制に影響を与えた例

REACH（2006）	化学物質の審査及び製造等の規制に関する法律改正（2009）
新車CO2排出規制（2009）	乗用車燃費規制企業平均燃費方式（2020）
GDPR（2016）	個人情報の保護に関する法律における追加安全措置（2018）

- **違反は事業存続や企業価値の毀損に直結**
- **提供者・販売者・使用者それぞれ責任を負う**

富士通の取り組み: AI倫理を重視

富士通AIコミットメント 2019

1. AIによってお客様と社会に価値を提供します
2. 人を中心に考えたAIを目指します
3. AIで持続可能な社会を目指します
4. 人の意思決定を尊重し、支援するAIを目指します
5. 企業の社会的責任として、AIの透明性と説明責任を重視します

全社AI倫理教育 2020

イベントを通じた発信

世界デジタルサミットで
時田社長がAI倫理の取組み発信
2022.06

AI4People設立 2018

Fujitsu Focuses on Social and Ethical Impact of AI with AI4People Forum

29th May 2018

4 PEOPLE

AI4People actively contribute to the debate on how the Commission implement the AI regulation

＊ 富士通(欧州研究所)は設立メンバー

Co-Creation & Collaboration: Focus on Ethics – AI4People Deep Dive

15th October 2019 | GA_Admin

Fujitsu Laboratories of Europe was a founding member of the AI4People initiative, and since then has been participating actively on behalf of Fujitsu in the EU AI Alliance organisations. Fujitsu was recently selected as one of just 50 members for an exclusive "Deep Dive" interview conducted by the AI4EU project – part of the EU's pilot process towards refining the EC's Ethics Guidelines for Trustworthy AI

AI倫理研究センター 2021

AI倫理ガバナンス室 2022

PRESS RELEASE

2022年1月28日
富士通株式会社

AIなど最先端テクノロジーの社会浸透・信頼確保の実現を目的とした、AI倫理ガバナンス室の新設

Fujitsu AI Ethics for Fairness

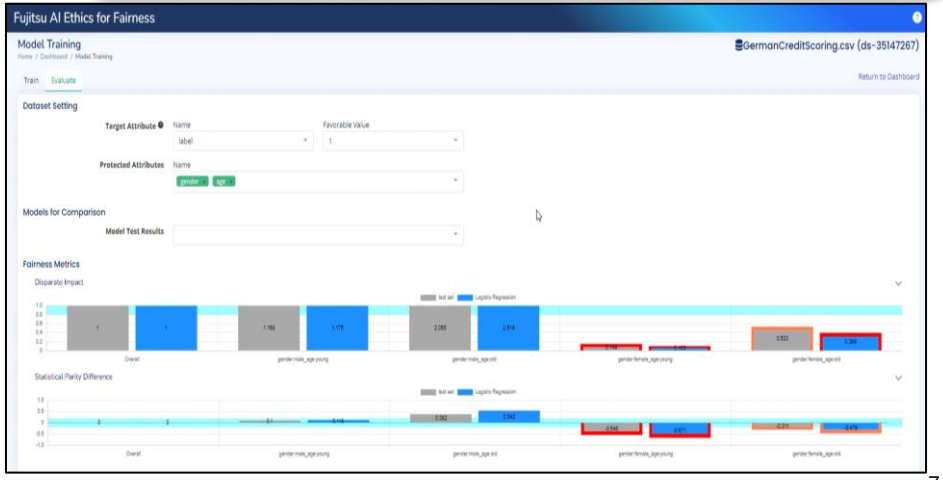
Fujitsu AI Ethics for Fairness

FUJITSU

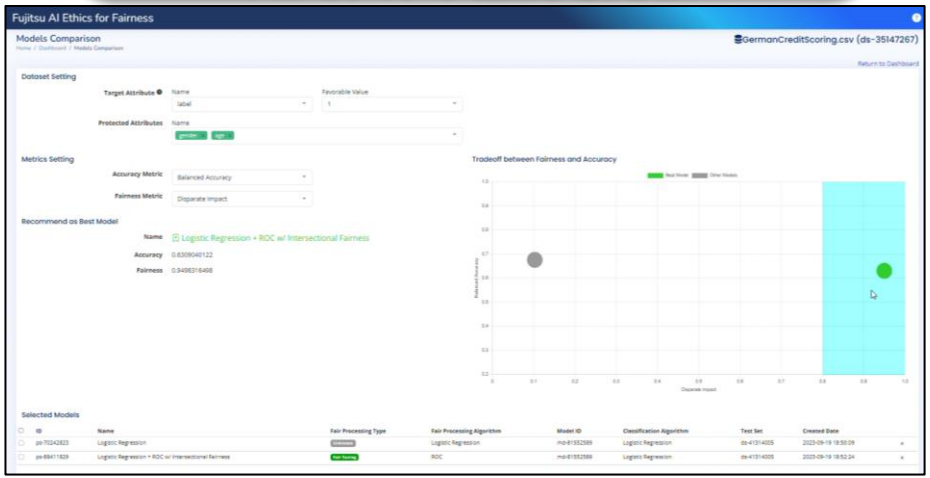
データの公平性の事前確認から、AIモデルの学習、構築したAIのバイアス緩和まで、実行可能

データのアップロード → データの公平性検証 → AIモデルの学習 → AIモデルの公平性改善 → AIモデルの比較

そのままデータを学習したAIモデルとバイアスを緩和したAIモデルを比較



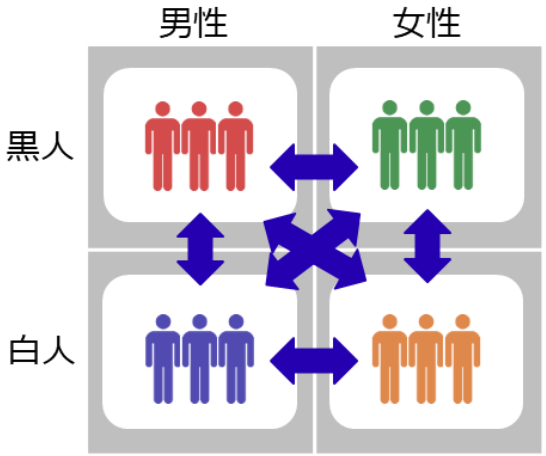
精度と公平性指標でAIモデルの性能を比較



Fujitsu AI Ethics for Fairness：特長と他社比較

独自の「交差バイアス緩和技術」が網羅的に検証することで、人間が気付きづらい AI判断のバイアス(偏り)も、検出・改善

交差バイアス緩和技術



特徴の組合せによって生じる
あらゆるグループ間のバイアスを軽減する
（例：黒人・女性、未婚のシングルマザー）

企業	単純なバイアス（1属性）検出・改善	気付きづらいバイアス（2属性以上）	
		検出	改善
FUJITSU	○	○	○
IBM OpenScale	○	×	×
Microsoft FairLearn	○	×	×
fiddler (Startup)	○	○	×

2. 交差バイアス緩和技術：実証実験

FUJITSU

○伊銀行のリアルデータから審査結果の偏りを検知し、AIモデルを改善

- 外国人女性の審査合格率が、単に外国人や女性の合格率よりも、大幅に低いことを検知

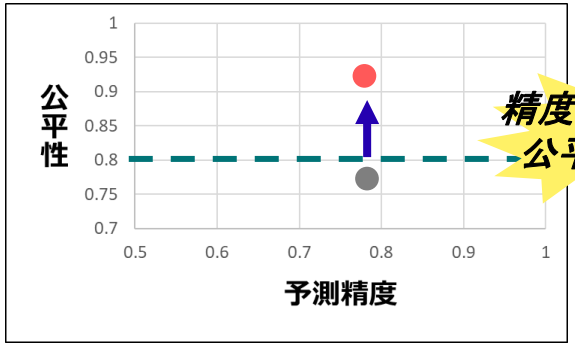
	（全体）	男性	女性
（全体）	－		
自国人			
外国人			

男性／女性では
ほとんど差なし

新たに
発見！

外国人x女性の
合格率が低い

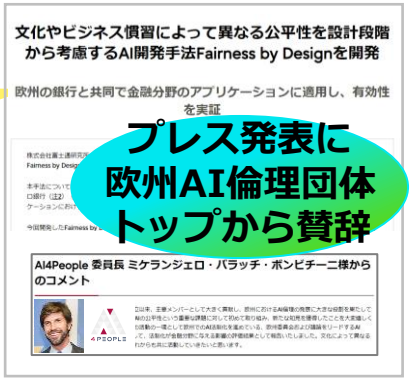
- 精度を保ちつつ、公平性を大幅に改善



精度劣化せず
公平性改善

交差バイアス緩和技術により実現

- プレス発表し、メディア・調査レポートで大きな反響



プレス発表に
欧州AI倫理団体
トップから賛辞

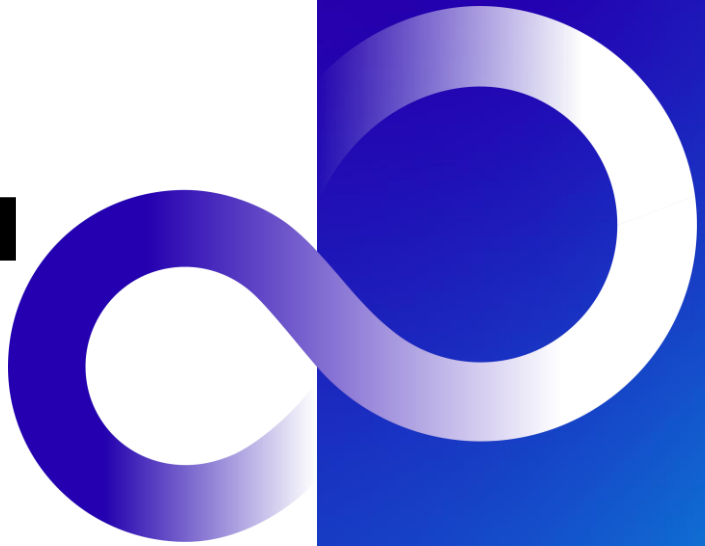
日経産業新聞に掲載

Gartnerレポートに掲載

注意事項

- 本サービスで検出した偏りや緩和結果は、現実に内在する社会的な差別や偏りを必ずしも正確に把握・緩和するものではありません
- 本サービスはすべてのバイアスの緩和を保証するものではありません
- 本サービスで作成されるモデルは必ずしも高い精度や公平性を保証するものではありません
- 本サービスで得られた結果について必ず利用者自身で偏りや、緩和の結果を確認してください

Thank you



FUJITSU

ご参考）公平性指標

FUJITSU

○ 技術的・法的には、**Disparate Impact（DI）**が標準的に使われる

$$\mathbf{DI} = \frac{\text{マイノリティ採用者数}}{\text{マイノリティ応募者総数}} \div \frac{\text{マジョリティ採用者数}}{\text{マジョリティ応募者総数}}$$

○ アメリカの雇用においては、もともとは1971年にState of California Fair Employment Practice Commission (FEPC) により、**DIが0.8以上**という基準が定められた。これは4/5ths rule (80% rule)と呼ばれており、法的な基準にもなっている。

○ アメリカの「雇用者選抜手続きに関する統一ガイドライン」によれば、差別訴訟事件において、「いかなる人種・民族・性別集団の採用理宇についても、それが最も高い採用率を示す集団と比べて5分の4（あるいは80パーセント）以下であるときは、一般に不利な影響の証拠となる」(*)

(*)「アメリカにおける人事採用政策に対する法的規制」(1991)
https://www.jstage.jst.go.jp/article/jaiop/5/1/5_19/_pdf/-char/ja