学術会議公開シンポジウム　2023年11月25日

「AI時代における哲学・美学・倫理学・宗教学」

- AI における公平性 -

理化学研究所 革新知能統合研究センター

上田 修功

Instruct-GPT

Training language models to follow instructions with human feedback

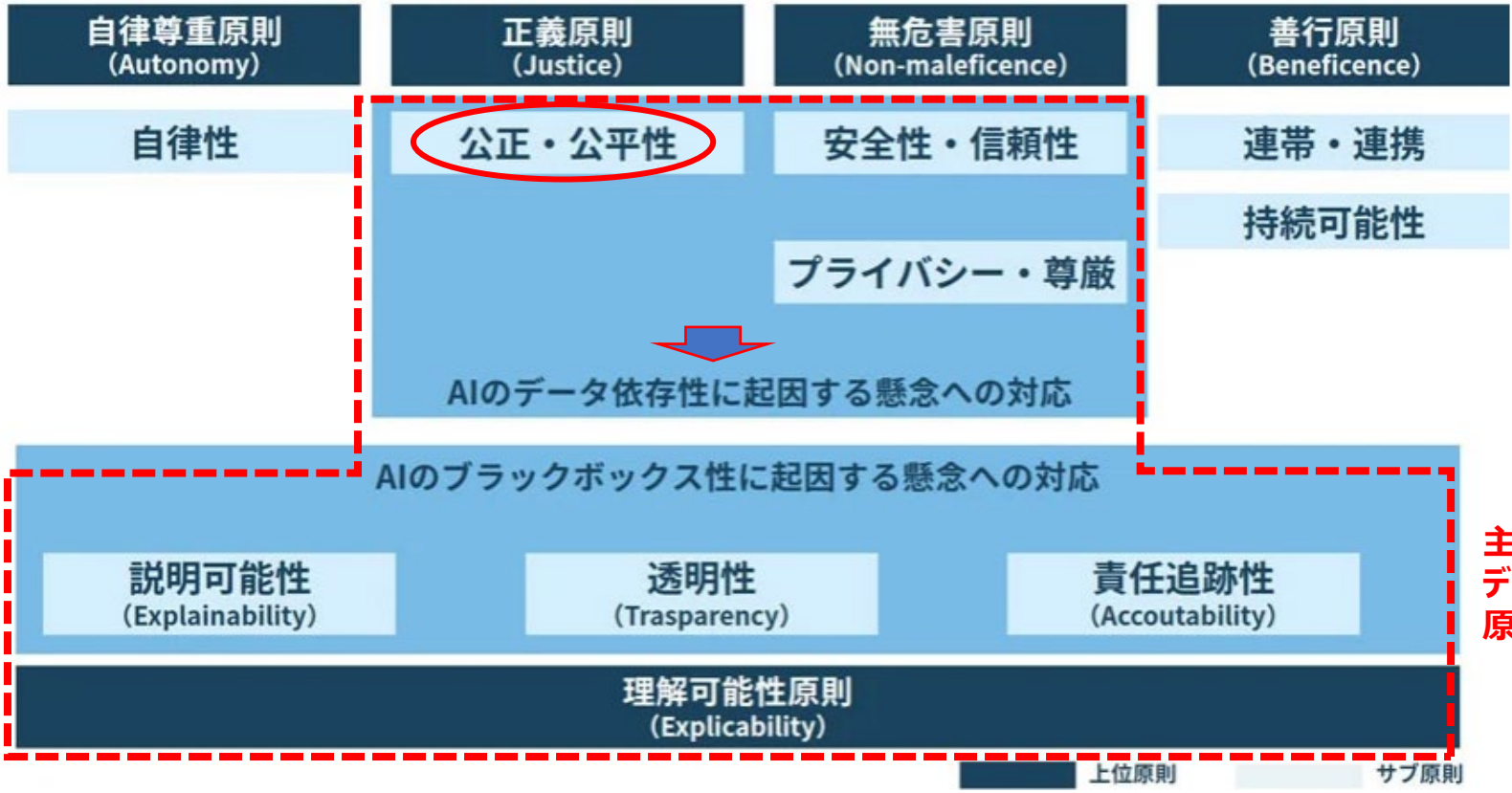
Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*
OpenAI

arXiv:2203.02155v1 [cs.CL] 4 Mar 2022

- 主に有害情報関係に言及
- 人手評価と自動評価によりGPT-3より改善

人間のフィードバックによる微調整が、言語モデルを
人間の意図に合わせるための有望な方向性である

世界各国の公的文書にみるAI５原則



主にシステムやデータに係る原則

参考文献：

A Unified Framework of Five Principles for AI in Society

https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/7

AI 原則とは？｜信頼できるAI のために世界各国が注目する基本理念

https://www.nri-secure.co.jp/blog/ai-principle

AIにおける(性)差別関連事例

- 機械翻訳で代名詞に性別の差がないトルコ語をhe/sheの差のある英語に翻訳すると，職業との共起に差が生じる。

https://medium.com/coinmonks/ai-doesnt-have-to-be-conscious-to-be-harmful-385d143bd311

- AmazonのAI（人工知能）を活用した人材採用システム（履歴書に書かれた約5万個のキーワードを抽出・分析）は、女性を差別するという機械学習面の欠陥が判明し、運用を取りやめる結果になった。

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

- 2019年UNESCO（ユネスコ：国連教育科学文化機関）は、AI音声アシスタントのデフォルトの声が女性となっていることはジェンダーの偏りを強めると主張する報告書を発表した。

- 顔認識ソフトウェアにおける人種差別: Googleフォトのソフトウェアは、写真に写っている物体や顔を認識して、人や動物などに基づいて分類し、写真を保存・整理するが、ある男性、ジャッキー・アルシネは、アフリカ系アメリカ人の友人の一人が写真の中でゴリラと表示されていた。

https://medium.com/coinmonks/ai-doesnt-have-to-be-conscious-to-be-harmful-385d143bd311

- アフリカ系アメリカ人の人名でGoogle検索すると、逮捕記録を示唆する広告が表示される。その広告には個人の逮捕記録をチェックできるサービスを提供するリンク(Instant Checkmate)が設定されていた。

https://queue.acm.org/detail.cfm?id=2460278

何故AIで(性)差別が生じるのか

機械学習タスクは、学習データ作成（特徴量の選定、データの収集）および、目的関数の設定からなるが、これらに(暗に)差別が混入する

➡ AIが差別を生み出すのではなく、むしろ社会の差別や偏見が差別的なAIを生み出している

例：SNSのAIチャボットが差別的な発言をする

➡ データ駆動型アプローチであるAI技術においては、個々人の（性）差別を解消しない限り、差別の根絶は困難

人間中心のAI社会原則

内閣府AI戦略

(6) 公平性、説明責任及び透明性の原則

「AI-Ready な社会」においては、AI の利用によって、人々が、その人の持つ背景によって不当な差別を受けたり、人間の尊厳に照らして不当な扱いを受けたりすることがないように、公平性及び透明性のある意思決定とその結果に対する説明責任（アカウンタビリティ）が適切に確保されると共に、技術に対する信頼性（Trust）が担保される必要がある。

- AI の設計思想の下において、人々がその人種、性別、国籍、年齢、政治的信念、宗教等の多様なバックグラウンドを理由に不当な差別をされることなく、全ての人々が公平に扱われなければならない。
- AI を利用しているという事実、AI に利用されるデータの取得方法や使用方法、AI の動作結果の適切性を担保する仕組みなど、用途や状況に応じた適切な説明が得られなければならない。
- 人々が AI の提案を理解して判断するために、AI の利用・採用・運用について、必要に応じて開かれた対話の場が適切に持たれなければならない。
- 上記の観点を担保し、AI を安心して社会で利活用するため、AI とそれを支えるデータないしアルゴリズムの信頼性（Trust）を確保する仕組みが構築されなければならない。

https://www8.cao.go.jp/cstp/ai/aigensoku.pdf

機械学習と公平性に関する声明

2019年12月10日

人工知能学会　倫理委員会
日本ソフトウェア科学会　機械学習高額研究会
電子情報通信学会　情報論的学習理論と機械学習研究会

1．機械学習は道具にすぎず人間の意思決定を補助するものであること

2．私たちは、公平性に寄与できる機械学習を研究し、社会に貢献できるよう取り組んでいること

関連シンポジウム：　機械学習と公平性に関するシンポジウム（2020年1月9日）

１．機械学習は道具にすぎません

機械学習はあくまでも道具にすぎず、その使い方を定めるのは人間です。機械学習は人類社会の繁栄に大きく貢献できる可能性を秘めているとともに、不適切な利用をすれば人類社会の利益に反する可能性もあります。機械学習は過去の事例に基づいて未来を予測しますから、偏りのある過去に基づいて予測する未来は、やはり偏りのあるものになりかねません。もし、過去と異なる「あるべき未来」を求めるのであれば、機械学習による予測や判断が公平性を欠くことがないように人間が機械学習に注意深く介入する必要があります。

同時に、「何が公平か」については、科学技術や工学だけの問題ではなく、現在の人類社会が何を求めているか、という価値観の問題抜きには語れません。機械学習という「道具」を正しく使うためには、それが「公平性」という私たち人類社会の価値観に対して、どのような影響を与えるかを正しく理解し、そのリスクを評価し、方策について合意しなければならないのです。この点は、私たちだけではなく、機械学習に携わる技術者や利用者、経営者、そして組織や社会の全体が把握し向き合っていく必要があります。

引用元：http://ai-elsi.org/wp-content/uploads/2019/12/20191210MLFairness.pdf

2. 私たちは機械学習で公平性に寄与します

私たちは、機械学習の利用が社会の不利益になってはならないと考え、この問題を解決するために、行動指針と技術開発の双方から真摯に取り組んでいます。IEEE Ethically Aligned Design では機械学習の不適切な利用ないしは誤用、悪用を戒め、その対策を具体的に記述しています[3]。人工知能学会では、自らの社会における責任を自覚し、社会と対話するために、学会会員の倫理的な価値判断の基礎となる倫理指針を 2017 年に定めました[4]。我が国社会の様々なステークホルダ（その一部は、私たちでした）が集まって、高度な情報技術を社会でどのように使っていくべきかを議論し、その結果が、内閣府「人間中心の AI 社会原則」として 2019 年 3 月に公開されました[5]。その基本理念の 1 つは多様性と包摂であり、高度な情報技術の利用にあたっては「公平性のある意思決定とその結果に対する説明責任」を担保するように求めています。

これらに呼応して、私たちも公平性の様々な側面をいかに定量的に評価し、実現していくかについての研究を進めています。最近の主要な研究集会では必ず機械学習の公平性に関する研究発表がありますし、世界的にも公平性に関する研究論文の数は増えています。実は「公平性とは何か」を機械学習の言葉で数理的に突き詰めていくと、多数のバリエーションがあることがわかります。人々が何を公平と考えるか、様々な基準を機械学習の言葉で表現しなおすことによって、「公平」という概念をより明確なものにしていくこともできるのです。このように、私たちは、機械学習によって公平性に起きうる問題を防ぐだけでなく、機械学習をきっかけとして公平性のあり方を定義、議論することにも真摯に取り組んでいます。

引用元：http://ai-elsi.org/wp-content/uploads/2019/12/20191210MLFairness.pdf

Fairness-aware Machine Learning

公平性を配慮した機械学習技術

2017年ごろより研究が加速（EUのGDPRの影響？）

Note: Bugbears or Legitimate Threats? (Social) Scientists' Criticisms of Machine Learning

各種のバイアス　Barocas, 2016

- **データバイアス**：　データ作成者の偏見や認知バイアスなどにより学習データに偏りが生じる
- **標本選択バイアス**：　予測対象の集団が学習データに含まれていないことによるバイアス
- **帰納バイアス**：　少数事例を例外、外れ値として扱うことによるバイアス（例：データの大半が男性データ）

Biased algorithms are easier to fix than biased people

Mullainathan 2019

公平性の形式的定義

S：モデル中のセンシティブ特徴
（性別，年齢，人種，宗教，政治指向 など）

S=1: 配慮が必要
S=0: 配慮が不要

X：その他の特徴

Y：Y=1: 不利な判定
Y=0: 有利な判定

$\hat{Y}$: AIによる予測値

データバイアスの解消

$\hat{Y} \perp\!\!\!\perp S$ （$\hat{Y}$とSが統計的に独立）

不利な判定と有利な判定がなされる割合を
Sと独立に同じとすることでバイアスを解消する

$P(\hat{Y}=1|X,S=1) = P(\hat{Y}=1|X,S=0)$
$P(\hat{Y}=0|X,S=1) = P(\hat{Y}=0|X,S=0)$

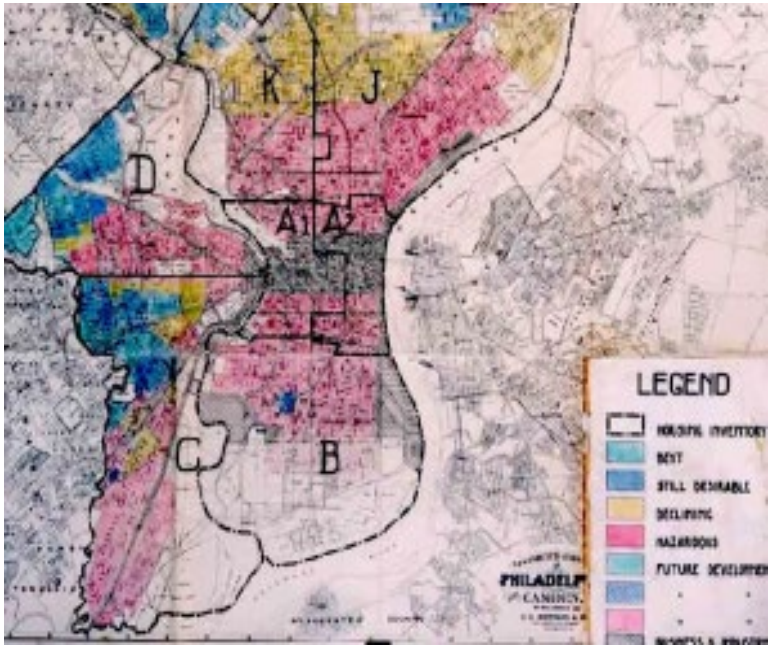
Note: Red-Lining Effect

Calders, 2010

ローン貸付の人種差別

住居情報が間接的に人種情報として
使われている

XとSの統計的独立性が重要



選択バイアスの解消

ランダムサンプリング，層別化

観測データセットが母集団の代表となるようにする

- 多用なデータを収集する
- 不均衡なデータセットを調整し，均等なサンプルを確保する

解釈可能性と透明性の向上(XAI)が重要

帰納バイアスの解消

$\hat{Y} \perp\!\!\!\perp S \mid Y$ （Yの値は正しいと仮定） Zafar, 2017

$P(\hat{Y}=1|Y=0,X,S=1) = P(\hat{Y}=1|Y=0,X,S=0)$ かつ

$P(\hat{Y}=1|Y=1,X,S=1) = P(\hat{Y}=1|Y=1,X,S=0)$

偽陽性率（$Y=0$を$\hat{Y}=1$と誤る率）と真陽性率（$Y=1$を$\hat{Y}=1$と正しく判定する率）をS=0,1の両者のケースで等しくする

Ex. 出所後の再犯予測において，白人は真陽性率が高く，
黒人は偽陽性率が高い

Angwin et al., 2016

集団公平性と個人公平性

集団公平性：　個々人は集団として公平に扱われる

- $P(Y \mid S=s) = P(Y)$ for all s in S

例：　男性と女性の各グループで平均的に取り扱いが等しければOK

個人公平性：　個々人が公平に扱われる

- $P(Y \mid S, X=x) = P(Y \mid X=x)$ for all x in X

Xが所与の下で、YとSが統計的に独立

例：性別以外の特徴が全て同じであれば、同一に扱われるべき

データマイニング，機械学習における公平性関連国際会議

- ICDM2012併設WS：
  Discriminant and Privacy-aware Data Mining
- NIPS2013, ICML2014 併設WS:
  Fairness, Accountability, and Transparency in Machine Learning
- NIPS2016シンポジウム：
  Machine Leaning and Law
- ICDM2016併設WS:
  Privacy and Discriminant in Data Mining
- KDD2019 チュートリアル:
  Fairness-aware Machine Learning: Practical Challenges and Lessons Learned
- ACM 2020
  Conference on Fairness, Accountability, and Transparency

まとめ

- AI(機械学習）は道具に過ぎず，データ収集，アルゴリズム設計，運用面で人間を介して差別が生まれ得る
- AIにおける公平性の議論は2017年代より活発化し，機械学習の国際会議でもWSやチュートリアル等で不公平性を防ぐ技術の研究が進行している
- 公平性とは何か，公平性の在り方についての議論が必要