障害者のための AI における公正を目指して：ひとつの研究ロードマップ

出典・凡例

本稿は，Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, Meredith Ringel Morris. (2020). Toward fairness in AI for people with disabilities SBG@a research roadmap. ACM SIGACCESS Accessibility and Computing, pp. 1 の要約である。

要約作成にあたり，原語を付すために()を用いた。

なお，disability および people with disabilities という語の訳にあたっては，日本において法律上用いられる用語であることから，それぞれ「障害」，「障害者」とした。また本論文は，日本において法律上も前提とされている，障害の社会モデルを前提としている（本論文注 1)）。

アブストラクト

AI テクノロジーは障害者 (people with disabilities) の生活に著しい影響を与える潜在性がある。じじつ，障害者の生活を改善することは，多くの最先端の AI システム――たとえば，耳の聞こえないあるいは聞くのが困難な人びとのために動画に字幕をつける自動音声認識や，発話ないし認知障害をもつ人びとのためにコミュニケーションを補強できる言語予測アルゴリズムなど――にとってのひとつの動機の原因 (motibator) である。しかしながら，ひろく用いられた AI システムは障害者にとって適切に働かないばかりか，彼らを積極的に差別する可能性すらある。障害者のための AI における公平性 (fairness) に関するこれらの考慮は，これまでほとんど注意を向けられてこなかった。本ポジションペーパーにおいてわれわれは，設計，開発，テストにおいて配慮がなされなかった場合に，いくつかの AI テクノロジーがどのように特定の障害をもつ者たちに影響を与えうるかという点について懸念のある潜在的な諸領域を特定する。われわれが目指すのは，さまざまなクラスの AI がさまざまなクラスの障害とどのように相互にかかわり合いうるかという点についてのこうしたリスク精査によって，データ収集，これらの仮説のテスト，より包摂的なアルゴリズムの構築に必要な将来の研究へのロードマップを提供することである。

本資料は，JST，RISTEX，JJPMJRX19B5 の支援を受けた研究成果の一部です。

キーワード

人工知能；機械学習；データ；障害；アクセシビリティ；包摂 (inclusion)；AI の公正 (fairness)；AI のバイアス；倫理的 AI

イントロダクション

- AI システムはますます現代の生活に浸透しており，そうであるからこそ全員にとって公正に機能することが重要な課題となっている。AI システムが公正に機能しているかを理解するには，評価のために包摂的なツールやプラクティスが生み出されたり，非包摂的なデータなどがシステム・トレーニングにおいて用いられていないかなどを見定めたりする必要があるところ，これまで，障害者にとっての AI の公正に関する検討は，ほとんど注意を払われてこなかった。
- 本稿でのリサーチ・アジェンダは次の諸点にある。(1) 障害者にとっての包摂性の問題が AI システムに影響を与える諸点を同定する；(2) 失敗した場合のシナリオと，既存のバイアス軽減技術がどの程度機能しているかを把握するために包摂性仮説 (inclusion hypotheses) を試す；(3) 複製と包摂を支えるためにベンチマークとなるデータセットを作成する；(4) 障害者に関する現状の手法によるあらゆる欠点を解決するため，新しいモデリング技術，バイアス軽減技術，誤差測定技術を開発する。
- 本論文においてわれわれはまず，現在の AI システムにおいて鍵となるクラスがさまざまなクラスの障害について特定の考慮を迫る可能性がどの程度あるかを検討する。さらに，そもそも AI における特定のカテゴリーを作ることが倫理的か否かという論点も重要なものであると考えており，以下で述べるさまざまな障害は，単に記述するものにすぎず，そうしたシステムが作られるべきであることは含意していない。

既存の AI システムの障害者にとってのリスク精査

- ここでは，既存の AI システムのクラスを連関する機能によって腑分けし，AI システムが問題となりうる対象となる障害をもつ者を同定している。もっとも，これは将来の研究の出発点にすぎず，網羅的なものではない。

コンピュータ・ヴィジョン

- コンピュータ・ヴィジョンは，スチルカメラやビデオカメラの入力を分析し，顔，身体，あるいは対象の存在 (presence) や属性 (attributes) といったパターンを特定する。

コンピュータ・ヴィジョンのアルゴリズムの公正性を設計しテストする場合，個人の身体的外見（顔の特徴，表情 (facial expressions)，身体の大きさや均整 (proportions)，補助器具の存在，典型とは異なる動きの属性）が重要な考慮要素となる。

顔認識

- 顔認識システムには，顔の存在を特定する機能，および／または顔の探知 (*detection*)，特定 (*identification*)，認証 (*verification*)，分析 (*analysis*) といった，属性に関する推定を行う機能が含まれる。
- われわれの仮定は，そうした技術は，トレーニング・データの収集およびモデル評価の際に考慮されない場合，顔の特徴および表情に違いをもつ人びとにとってうまく機能しない可能性がある，というものである。
- たとえば，顔分析ソフトのさまざまな側面は，ダウン症，軟骨無形成症，口唇裂といった状態にある人びと，視覚障害により目以外にも違いが生じている人びと，あるいは視覚障害によりサングラス (dark glasses) 等を身につける必要のある人びとなどにとって，うまく機能しない可能性がある。

身体認識

- 身体認識システムには，身体の存在を特定する機能，および／または身体の探知，特定，認証，分析といった，属性に関する推定を行う機能が含まれる。
- 身体認識システムは，身体の形質，姿勢，あるいは動きの違いといった特徴をもつ障害者にとってうまく機能しない可能性がある。
- たとえば，ALS や四肢麻痺により運動の範囲がきわめて制約されている人びとは，身体認識システムが相互作用によってのみ認識されていればその使用から締め出される。また，自動運転車の歩行者探知アルゴリズムは，脳性小児麻痺やパーキンソン病の人びと，高齢者，車椅子利用者といった姿勢に違いをもつ人びとの例を含んでいない。

対象認識，シーン認識，テキスト認識

- 対象認識，シーン認識，光学的文字認識 (OCR) システムは，共通の対象，ことば，テキスト，筆跡等，および出力ラベル，キャプション，（位置，活動，関係といった）プロパティを認識する。
- 写真から対象を認識するシステムのほとんどは，目の見える（しばしば地理的にも収入的にも恵まれた）人びとの，SNS で撮られた写真をデータセットとして利用しているが，目の見えない人びとの写真は，フレーミングや光やアングル等の点で質が大きく異なることから，画像処理プロセスでの誤差率が上昇しがちであ

る。こうした問題は，運動機能に障害をもつ人びとが撮った写真や筆跡についてもあてはまる。

スピーチ・システム

- 本論文で「スピーチ・システム」とは，発話の内容（つまり，ことば）や属性を認識する，あるいは記号によるインプットから発話を生成する AI システムを指す。発話の内容や明瞭性に影響を与えうる障害によって，スピーチ・システムの正確性と有用性を減ずる可能性がある。

音声認識(Speech Recognition)

- 自動音声認識システムは音声を取り入れ，テキストを出力するため，聴覚に障害がある人びとなどにとって重要なツールとなる可能性がある。
- もっとも，こうしたシステムは，発話が典型的でない人びと (people with atypical speech)（たとえば，女性，一部の高齢者，発話障害をもつ人びとなど）にとって正しく機能しない可能性がある。

音声生成

- 音声生成には，記号的なインプットから現実の音声を生成するテキスト・トゥ・スピーチ (text to speech) 等が含まれる。
- 理解できる発話速度のシステムデフォルトがどのようなものであるかは，具体的な障害の部分によって調整を行う必要がある。たとえば，認識障害または知的障害をもつ人びとにはよりゆっくりとした発話速度が必要となる可能性がある一方，視覚障害をもつ人びとはそれでは遅すぎると感じる可能性がある。

発話者分析

- 発話者分析には，発話者の特定，発話者の認証，および発話者の属性に関する推定を行う機能が含まれる。
- ユーザーの個人的な特徴（ジェンダーや年齢など）について推定を行う発話者認識と発話者分析は，たとえば構音障害をもつ人びとなどの障害者にとってうまく機能しない可能性がある。

テキスト処理

- テキスト処理システムは，テキストデータの内容を理解することに関わる機能を果たす。こうしたシステムは，認知障害や知的障害をもつ人びとにとって正確性や公正性の課題をもつ可能性がある。

テキスト分析

- テキスト分析は，テキストをインプットとして取り込み，内容の属性や著者の属性を探知しようとする。こうしたシステムは，読み書きの能力について障害をもつ人びとや認知の相違をもつ人びとに役立つ可能性がある。
- 認知障害や知的障害は，テキスト分析システムの多くの側面における効率や有用性に影響を与える可能性がある。このことから，テキスト分析システムは認知障害や知的障害をもつ人びとにとって正確性や公正性の課題をもつ可能性がある。

統合的 AI

- 上記のシステムはいずれも単一のモデルを用いていたが，多くの複合的な AI システムは，複数のモデルを統合するアーキテクチャであり，より複雑な振る舞いをすることができる。

情報検索 (Information Retrieval)

- ウェブの検索エンジンを駆動するもの等の情報検索ツールは，さまざまな目的のために AI に依存しており，またインプットおよびアウトプットはさまざまなフォーマットでありうる。
- 情報検索システムは，障害者に対する既存のバイアスを負の方向に増大させる可能性がある。コンテンツ・ベースおよび行動ベースの広告のための AI システムはソーシャル・メディアだけでなく商業的情報検索システムの鍵となるコンポーネントであるが，広告アルゴリズムは商品やサービスに対して差別的価格づけ (differential pricing) を行うこと，あるいは雇用や賃貸等の機会において差別的表示 (differential exposure) を行うことなどにより差別的行動を積極的に広めてしまう可能性がある。情報検索システムは，とりわけ認知障害または知的障害をもつ人びとにとって障壁となる可能性がある。

会話エージェント

- 会話エージェントは，教育やカスタマー・サービス等のさまざまな実践的場面のためにエンド・ユーザーに会話の機会を提供する。このエージェントはテキスト分析や発話者分析等のさまざまなモデルを利用することで，知的障害をもつ人びと等にとって認知機能の援助となる可能性がある。
- 注意深く作られない場合，会話エージェントは，さまざまな会話におけるステレオタイプ的内容を返すことなどにより，障害者に対する既存のバイアスを増大させる可能性がある。また，認知障害または知的障害をもつ人びとにとってうまく機能しない可能性もある。多様な認知能力および知的能力の人びとからのデータを含むコーパスにより会話エージェントを訓練することが，とくに重要である。

その他の AI を用いた技術

- 外れ値探知，統計測定 (aggregate metrics) による評価システムの実践，目的関数の定義，および真のユースケースや現実世界にある複雑性を捉えないトレーニングデータの使用といった，AI システムの構成要素となる多くの技術および実践が障害者に対するバイアスにつながる可能性があるという点も，検討に値する。

議論

・不公正な AI により生じる潜在的な害の類型としては，以下のものが挙げられる。

①サービスのクオリティ；発話障害をもつ人びとのインプットを認識できないスマートスピーカーなど

②配分における害；自閉症の人の精神状態ないしパーソナリティについて誤った予測をインプットとして自動雇用システムに用いることなど

③中傷ないし軽視(denigration)；障害者からのインプットを無効な外れ値として誤ったフラギングをすることなど

④ステレオタイプ化と過大または過少代表；検索結果においてステレオタイプ化された，またはうまく代表されていない内容を返すことで障害者に対する既存のバイアスを負の方向に増大させることなど

- ①②④についてはベンチマーキングによる測定の客観的公正性を測ることでバイアスを明らかにするのに足りる可能性もあるが，③およびステレオタイプ化についてはさらなる質的調査が必要となる可能性がある。
- 本論文が取り組んだのは，イントロダクションで掲げた 4 つのアジェンダのうち第一のもののみであり，残る 3 つは今後の課題である。なかでも，特定のユーザー・グループにパーソナライズされたモデルを作るのではなく幅広い人口層を通して公正な一般的モデルを作ることがどの程度可能か（あるいは望ましいか）という点を考えることが重要な課題となる。障害の「ロングテール」性に鑑みると，パーソナライズすることの必要は高いといえる。もっとも，パーソナライズされたモデルを訓練する必要性によって，AI モデルがデフォルトで機能する人びととそうでない人びととの 2 階のシステムが生じるだけでなく，障害者にとってさらなる障壁が顕在化する可能性もある。
- 障害者が AI システムの評価においてだけでなく，有意義な利用シナリオ，誤差測定，および諸政策の特定の場面にも関わることが，公正な AI の開発にとって不可欠である。

結論

・ 究極的には，われわれの目標は，AI システムが障害者に利益を与えうる巨大な可能性を実現しつつ，本論文で概説したような潜在的な危険を避けることに役立つような新設計のガイドライン，アルゴリズム技術，誤差測定を生み出す点にある。

（松本　有平）