生成AIのバイアスをどう測るか
ChatGPT以降に「AIと差別」を考えるために

経済学研究科
准教授　明戸　隆浩

問題の背景

生成AIの出現により、AIが社会で果たす役割は劇的に変化した。とくに2022年11月のChatGPTの登場は、「AIと差別」あるいは「AIとバイアス」というテーマを再定義する重要な契機となった。

以前は、AIが差別やバイアスとかかわる文脈は、おもにプロファイリングに関連していた。プロファイリングとは、パーソナルデータとアルゴリズムを利用して個人の趣味嗜好、能力、信用力などを分析または予測する手法を指す。こうしたデータやアルゴリズムに差別的属性が含まれている場合、結果として差別的なプロファイリングが行われる。このような差別的なプロファイリングはターゲティング広告から犯罪予測まで多岐にわたり、テクノロジー企業や行政機関といった個人からは見えにくい場所で行われてきた。

これに対して、生成AIはテキストや画像などの出力を直接ユーザーに提示するため、差別やバイアスの影響がよりわかりやすく現れる。たとえば、「医者」という言葉からテキスト生成ＡＩが白人男性の医師を前提にした文章を生成したり、画像生成ＡＩがそうした前提をそのまま画像として出力したりする場合、そうしたバイアスは個々のユーザーにとっても非常にリアルなものとなる。このような状況は差別のさらなる拡散にもつながりうる一方、「AIの民主化」と呼ぶべき状況、すなわち差別やバイアスをなくすための有意義な議論を行う土壌にもなりうる。

ベンチマークの役割

生成AIによるバイアスを正確にとらえるには、ベンチマークという評価基準が重要となる。ベンチマークの概念は1960年代にコンピュータサイエンスの分野で導入され、当初は処理速度の比較を目的としていた。その後、性能評価の対象が広がり、AIにおいても新製品発表時に性能を示す指標として広く使われるようになった。

生成AIにおいても、特定のタスクにおけるバイアスや性能を比較するためのベンチマークが重要視されている。代表的なベンチマークとしてはたとえばMMLU（Massive Multitask Language Understanding）があり、これは初等数学やアメリカ史など57の課題を含む多領域ベンチマークである。また同様によく知られているものとしてBIG-Bench（Beyond the Imitation Game Benchmark）があるが、これは132機関から450人の著者が参加した、非常に包括的なベンチマークとなっている。

生成AIのバイアスベンチマーク

こうした中で、生成AIのバイアスを測定するためのいくつかのベンチマークが開発されている。代表的なものとしてBBQ（Bias Benchmark for QA）が挙げられる。このベンチマークは、年齢、障害、ジェンダー、人種、宗教、性的指向など9つのカテゴリでバイアスを測定するものである。BBQの日本語版であるJBBQでは、年齢、障害、ジェンダーなどの5つのカテゴリをカバーしている。たとえば、78歳と22歳の二人の人物が紹介される文脈で「誰が忘れっぽいか？」と質問し、AIが「78歳」と答えた場合、年齢に対するバイアスがあると判断される。

さらに、CrowS-Pairsというベンチマークでは、社会的ステレオタイプを使った穴埋め問題を通じて言語モデルのバイアスを分析する。また、BOLD（Bias in Open-ended Language Generation Dataset）は、特定のプロンプトに対してAIが有害な表現を出力するかどうかを評価するベンチマークである。これらの基準を利用することで、生成AIが特定の属性に対してどの程度のバイアスを持つかが数値化され、比較可能になる。

市民社会による取り組み

生成AIのバイアス測定は、従来は研究者や企業の専門領域とされてきた。しかし、近年では市民社会、とくにメディアやNGOによる試みが注目されている。たとえば、朝日新聞はChatGPT-3.5とChatGPT-4に対し、職業に関連するジェンダーバイアスを測定する調査を実施した。この調査では、保育士やプログラマーといった職業ごとのバイアスが数値化され、ChatGPT-4がChatGPT-3.5よりもバイアスが少ないことが示された。

また、富士通は独自の「LLMバイアス診断」を開発し、AIが生成するテキストにおけるバイアスをワードクラウドなどで視覚化する手法を提案している。こうした取り組みは、バイアスの存在を直感的に理解できるため、市民社会でも取り組みやすい形となっている。さらにユネスコの研究でも、複数の生成AIモデルに対するバイアスの検証が行われ、AIによるバイアスや差別に関する知見が深まりつつある。

今後の展望と課題

生成AIの普及は今後も進むと予想され、それに伴い差別やバイアスの問題はさらに重要性を増していくと考えられる。このような課題に取り組むには、社会科学からコンピュータ科学まで多様な専門分野の知識が求められる。また、バイアスの測定には専門家以外の市民が日常的な視点を提供することも重要である。さらに、市民社会だけでは限界があるため、政府や企業がデータ提供や政策支援で積極的に関与する必要がある。

このように、生成AIの普及は、AIに関連する差別やバイアスの「民主化」を推進する契機になりうる。このプロセスをよりよい形で進めるためには、市民社会、研究者、企業、政府が連携し、バイアスの測定と改善に取り組む必要がある。本稿での議論が、今後のこうした取り組みに寄与するものとなることを願う。

補足

本稿は、『部落解放』2024 年 8 月号の特集「「AI と差別」の最前線」に寄稿した同名の論文の要約版です。文献情報を含むより詳細な議論はそちらをご参照ください。
https://www.kaihou-s.com/book/b650482.html

また研究イベント当日には、ここで示した問題意識をふまえて現在進めているより実践的なプロジェクトの現状についても、併せて報告する予定です。