日立財団
グローバル ソサエティ レビュー

〈論文〉

生成 AI が多文化共生社会に与える影響

Impact of Generative AI on Multicultural Society

岸本　充生 *1,2
Atsuo Kishimoto

AI の仕組みを，予測・認識・分類 AI と生成 AI に分けて説明し，各々が抱える倫理的・法的・社会的課題（ELSI）を整理する。バイオメトリクス（生体認証技術）の課題に触れたうえで，テキスト生成 AI や画像生成 AI における文化的バイアスの実例を示し，それらが差別やステレオタイプを増幅する可能性を指摘する。最後に対応策を検討したうえで，ユーザーが生成 AI の仕組みと限界を理解して利用することの重要性を指摘する。

This paper categorizes AI mechanisms into predictive/recognition/classification AI and generative AI, and examines the ethical, legal, and social issues（ELSI）associated with each. It addresses challenges in biometric technology and highlights examples of cultural bias in text- and image-generating AI, emphasizing their potential to reinforce discrimination and stereotypes. Finally, the paper explores countermeasures and underscores the importance of users understanding the mechanisms and limitations of generative AI to use it responsibly.

1.　はじめに―AI の時代

近年，人工知能（AI）の発展のスピードが増している。背景にあるのが，デジタル化によるデータ量の爆発的な増加と，コンピューターの計算能力の増強，そして機械学習の発展であり，これらが合わさることにより，説明文付きの大量のデータ，すなわちラベルのついた学習データセットさえ整えば，予測・認識・分類を目的とする AI モデルを開発することが可能となった。ウェブ閲覧履歴や購買履歴から，次に購入しそうなものを予測して個人ごとに最適化した広告を配信したり，オンライン市場でレコメンド機能を付与したりすることは日常的に行われている。顔認証技術はすでにスマホやパソコンを開いたり，建物や部屋への入場パスとしても利用されたりしている。複数のパーソナルデータを組み合わせて，新しい情報を生み出すことは，データによるプロファイリングとも呼ばれ，個人の内面や疾患など，時にセンシティブな情報を予測したりできるため，使い方によってはプライバシーの侵害や差別につながる恐れがあり，また，誤認識や誤分類があることも知られている。これらは新しい技術の社会実装に伴う倫理的・法的・社会的課題（Ethical, Legal and Social Issues: ELSI）と呼ばれる［1］。

2022年11月末に OpenAI 社からテキスト生成 AI である ChatGPT が発表されてからは，プロンプトで指示を入力するだけでテキストや画像などを生成してくれる生成 AI サービスが次々と生み出され，各社が競い合うように新しいモデルを発表している。生成 AI はインターネット上の莫大なテキストや画像を学習データセットとして利用することで成り立っており，個人情報保護や著作権の課題などを置き去りにして技術が先走っているような状態である。それ以外にも悪用された場合のリスクや誤情報を生み出す（ハルシネーションと呼ばれたりする）問題など，様々な ELSI が指摘されているものの，企業による開発競争がある種の国際競争や経済安全保障の側面も持つことから，規制の議論は後追いにならざるを得ない。こうした汎用目的 AI（General Purpose AI）とも呼ばれる

*1 大阪大学 D3 センター（旧・データビリティフロンティア機構）
The University of Osaka, D3 Center
*2 大阪大学社会技術共創研究センター（ELSI センター）
The University of Osaka, Research Center on Ethical, Legal and Social Issues（ELSI Center）

高度な AI については人間の制御が不能になるリスクなども指摘されている［2］。

本稿では AI，特に生成 AI の持つリスクの中から，文化的なバイアスを取り上げ，多文化共生社会にとって脅威となりかねないことを指摘し，対応策を考察するものである。第 2 節では予測・認識・分類 AI と生成 AI に分けてそれらの技術的な仕組みとすでに指摘されている様々な課題をまとめたうえで，法規制のグローバルな動向を簡単に記した。第 3 節では予測・認識・分類 AI と生成 AI をつなぐものとして，バイオメトリクス（生体認証技術）を取り上げた。第 4 節では生成 AI の持つ文化的バイアスに焦点を当て，いくつかの事例も交えながら掘り下げる。第 5 節ではそれらのバイアスに対抗する方法を述べる。

2.　AI の仕組みと課題

1）　予測・認識・分類 AI の仕組みと課題

本稿では生成 AI に対して，それ以前からある AI を，予測・認識・分類 AI と表記して区別する。図 1 に予測・認識・分類 AI の簡素化した作成プロセスを示す。上段が学習プロセス，下段が推論プロセスである。猫の画像に猫と説明ラベルを付けるように，大量のデータセットにラベルを付けて機械学習させることで学習済みモデルが構築される。そこに生データをインプットすることで予測・認識・分類といった機能が発揮される。

予測・認識・分類 AI の倫理的・法的・社会的課題（ELSI）は大きく分けて，学習データの取得部分，アルゴリズム部分，結果の使い方の部分で生じうる。学習データの取得においては，パーソナルデータを適正に取得したかどうか，学習データセットに偏りがあるかどうか，どういう学習データを使っているかが公開されているかといった点が指摘される。アルゴリズムについては，バイアスの問題が指摘されている。結果の使い方については，AI の出力データのみに基づいた自動的な意思決定は，場合によっては人間の自律性や尊厳を損なうものであるとされる。EU の一般データ保護規則（GDPR）では，プロファイリングに基づく自動化された意思決定に対して，異議を申し立てる権利，決定に服さない権利や知る権利などが規定されている。データによるプロファイリングについては，日本国内には規制はないが，自主的な取り組みのためのガイドラインが公開されている［3］。また，AI を活用する企業の多くは，AI 倫理原則や AI 倫理指針を策定しており，それらに基づいて自主的な取り組みが広く行われている［4］。

2）　生成 AI の仕組みと課題

生成 AI の仕組みも図 2 に示すように，基本的には予測・認識・分類 AI（図 1）と同じであるが，下段が生成プロセスとなる。テキスト生成 AI も画像生成 AI も学習データセットは，主にインターネット上からウェブクローラーによりスクレイピングされた膨大なデータからなる。有害なデータを取り除く（フィルタリング）などしてモデルが設計される。プロンプトとして指示を入力することで出力が得られる。

生成 AI の ELSI は多岐にわたる［1］。図 2 のプロセスに沿って挙げる。学習データの取得においては，著作権との関係，個人情報を含みうること，データの偏り，透明性の欠如，ただ乗り批判が挙げられる。フィルタリングにおいては途上国の労働者に，精神的負担の多い有害情報の削除やラベリングを低賃金で委託していたことも問題視されている。モデル設計においては，エネルギーや資源の消費量が多いこと，一部のテック系企業による寡占状態にあること，アルゴリズムのブラックボックス性などが指摘されている。プ

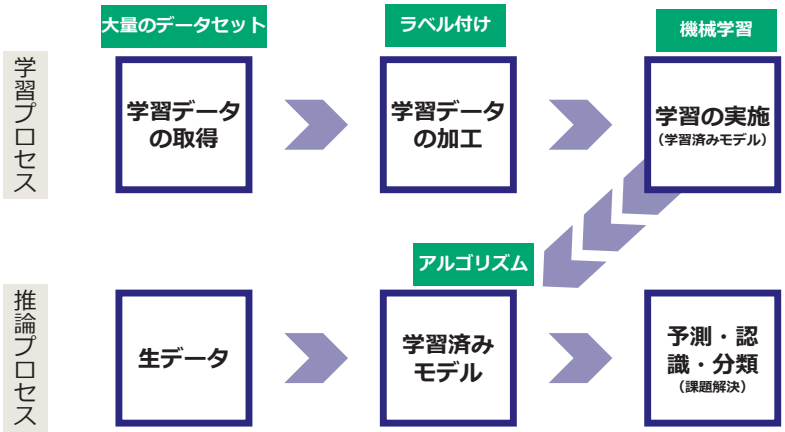

図 1　予測・認識・分類 AI の簡素化した作成プロセス（出典［1］図 1 を改訂）

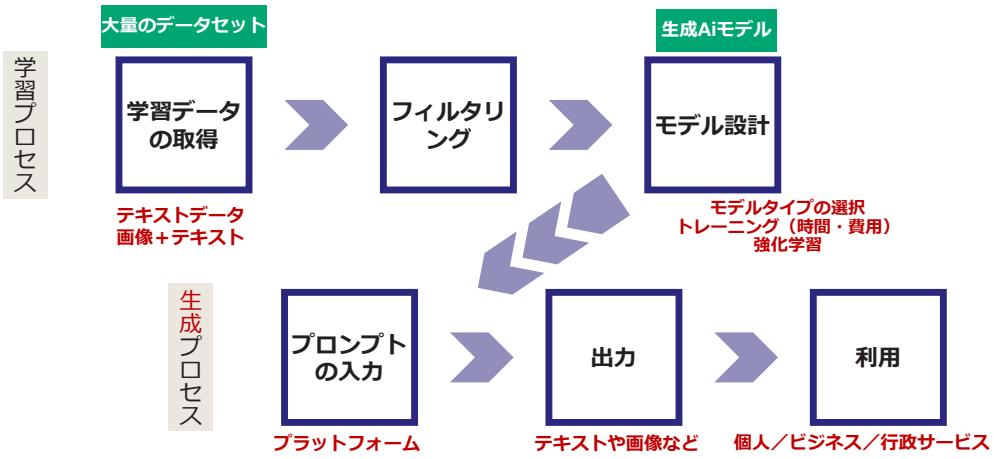

図 2　生成 AI の簡素化した作成プロセス（出典［1］図 2 を引用）

ロンプトの入力においては，著作権の侵害，個人情報や機密情報の漏洩，脱獄プロンプトと呼ばれる攻撃の可能性などが指摘されている。モデル出力については誤情報に加えて，バイアスや文化的周縁化が指摘されているがこれらは後述する。出力の利用段階では，著作権の侵害，擬人化による依存，偽情報の流布，差別的な使われ方，軍事利用や悪用の可能性などが指摘されている。

3）AI に対する法規制の動向

AI に対する法規制は欧州が先行している。EU では 2024年 8 月に AI 法が発効し，世界初の包括的な AI 規制枠組みが提案された［5］。守りたいものとして，健康・安全に加えて，基本的権利が挙げられ，法の支配，民主主義，環境も含めることが明記された。AI システムはそれらの「守りたいもの」へのリスクレベルに応じて，容認できないために禁止される AI 実践，高リスク AI システム，透明性が求められる特定の AI システム，それら以外の AI システムの 4 段階に分類され，さらに汎用目的 AI モデルと，それらの中でも大規模な影響を持つ（これを「システミックリスクを伴う」という）汎用目的 AI モデルについて，プロバイダ（提供者）やデプロイヤ（実装者）への取り組みが求められている。2025年 2 月からは，禁止される AI 実践に対する規制が先行して施行された。高リスク AI システムについては，実践規範（Code of Practice）が準備されている。米国では，連邦レベルでの法規制はないものの，国立標準技術研究所（NIST）が AI リスクマネジメント枠組みを策定し，また連邦取引委員会（FTC）などの連邦省庁による既存の法律権限の範囲内での対応がなされている。また国際的には OECD（経済協力開発機構）や G7（主要 7 か国首脳会議）の枠組みなどを利用した取り組みも進んでいる。2023年には G7議長国として日本が「広島 AI プロセス」を立ち上げ，「国際指針」と「国際行動規範」を取りまとめた。2024年末には，AI 開発企業等が「国際行動規範」の遵守状況を，質問票への回答という形で公表するためのページが OECD のウェブサイト上に公開されている。

3.　バイオメトリクス（生体認証技術）の多文化共生への課題

予測・認識・分類 AI の中でも，指紋，顔，虹彩などを利用するバイオメトリクス（生体認証技術），特に顔認識技術は，深層学習の発展とともに近年，性能の向上と利用が急速に進み，世界各地で議論を引き起こした［6］。生成 AI と同様，顔認識技術にも，インターネット上から勝手に収集した大量の顔写真を学習データセットとして用いて訓練されたモデルが用いられていたことから，批判を受けて，研究開発用として公開されていたデータセットの取り下げが相次ぐなどの混乱が生じた［7］。顔認識技術は米国や英国において当初，女性や有色人種への誤認識率の高さも社会問題となった。これは学習データセットの中身が白人男性に偏っていたことに起因している。米国ではブラック・ライブズ・マターの運動と時期的に重なったこともあり，サンフランシスコ市などの一部の地方自治体では公的機関が顔認識技術を使うことを禁じる条例が成立した。米国でも英国でも警察が，公共の場に置かれたカメラ画像に移った顔写真を，ブラックリストである容疑者等の顔データベースと照合するという形で，バイアスのある顔認識 AI カメラを利用したため，マイノリティへの差別問題として課題が表面化した。

また，国際人道援助の文脈においても，援助が（中抜きされずに）個人に直接届いたり，何らかの理由で ID を持たない人でも銀行口座が作れたりすることか

ら，顔認識技術を含むバイオメトリクスが利用されている。しかし，有力な国際援助団体の 1 つである Oxfam は2015年，バイオメトリクスの利用に関して，2 年間の自主的なモラトリアム（一時停止）期間を設けることを決めた。その終了に合わせて，バイオメトリクスが国際人道援助において将来的にどのように位置づけられるべきかに関して研究委託を行い，リスクとベネフィットを検討した報告書が2018年に公表された［8］。報告書では，リスクとして以下の 5 点が挙げられた。

R1：間違って適合したり，しなかったりする可能性（reliability）
R2：他者に利用されたり，意図しない目的で利用されたりする可能性（reusability）
R3：生体データの窃盗，喪失，悪用のリスク（security）
R4：データ漏洩が起こった際や間違った噂が広がることによる評判リスク（reputation）
R5：何らかの理由で拒否した場合に起こる排除の可能性（societal impacts）

その後も検討が続けられた結果，2021年には安全で責任あるデータの利用のために 7 つの原則や各種ポリシー，チェックリストなどを定めたうえで，バイオメトリクスを慎重に利用していく方針が定められるに至った［9］。

4. 生成 AI の文化バイアス

1） テキスト生成 AI のバイアス

テキスト生成 AI は大規模言語モデル（LLM）と呼ばれ，インターネット上のテキストや書籍情報などを学習データセットしていることから分かるように，社会が持っているステレオタイプや差別的な表現，攻撃的な言動など，現存する社会的バイアスをそのまま受け継ぎ，さらに最もありそうな出力を行うために，そうしたバイアスを増幅する性質を持っている。また，インターネット上のテキスト自体の偏りがあることも忘れてはならず，英語の占める割合が最も多く，欧米に関する情報は詳しいが，マイナーな言語や文化に関する情報が少ないことは容易に予想できる。

実際に，生成 AI モデルに文化的なバイアスを見出した研究も多数報告されている。アップル社の機械学習チームは，2023年初頭に発表された 4 つの LLM でテストをして，すべてのモデルが，様々な職業について，男性と女性に関して，技術職や管理職の文脈で男性を，ケア職や補助職で女性を過剰に関連づけるなど，米国労働局の統計に基づく統計事実よりもさらに偏ったステレオタイプを表現していることを見出した［10］。LLM は不均衡なデータセットで訓練されているため，人間のフィードバックによる強化学習があっても，不均衡を反映し，さらには増幅しさえする傾向があると評価した。

2024年の UNESCO（国連教育科学文化機関）の調査では，LLM の社会的バイアスを検出するための確立された方法として，異なる概念同士をどのように関連づけるかを測定する方法と，与えられたテーマに沿ってどのようなテキストを即興で生成するかを観察する方法が使われた［11］。3 種類の LLM を対象とした実験では，前者についてはおおむね，女性の名前は「家庭」，「家族」，「子供」，「結婚」と関連し，男性の名前は「ビジネス」，「重役」，「給料」，「キャリア」と関連した。後者については，完成していない文章を LLM に完成させるという実験が行われ，同性愛や特定の人種に対してネガティブな反応をする LLM が存在することを見出した。また，代表的でない文化やグループに対してよりステレオタイプに頼ることが多いことも明らかになった。

2） 画像生成 AI のバイアス

ソーシャルメディアの普及でますます人々はテキストよりも画像から情報を得る比率が高まっており，画像生成 AI によって引き起こされるバイアスの拡大再生産の問題にはより注意が払われるべきかもしれない。画像生成 AI モデルの多くはドイツの非営利団体である LAION が作成した LAION-5B と呼ばれる58.5億の画像 - テキストのペアからなるオープンなデータセットを学習データセットのベースとして利用している。これらはすべて誰でもアクセスできるインターネット上から収集されたものである。そのため，有害な画像や医療画像などのセンシティブな画像，また著作権で保護されたコンテンツを当然含むし，データ化されたコンテンツしか参照しておらず，インターネットの持つ文化的な偏りやステレオタイプをそのまま保持していることになる。生成 AI は（インターネットの世界で）最も「ありそうなもの」を生成するため，大統領や政治家，医者や社長などの画像を指示すると当初は白人男性の画像ばかりを生成していた。現実世界で男女比率がたとえ 7 対 3 であっても，生成 AI の仕組み上，生成される際には10対 0 になるのである。その後，各社はアルゴリズムにダイバーシティ配慮を組み込んだ様子が見られるが，以下に示すようにバイアスは根強いものがあるうえに，誰もが納得する「正解」がない課題でもある。

スタンフォード大学の Bianchi らは，誰でもインターネット上で使える画像生成 AI を使って，単純な特徴や職業，物体を含むプロンプトが，人種・性別・国籍に関する固定観念を強化することを見出した［12］。例えば「魅力的な人物」は白人的な特徴を持ち，「テロリスト」は黒髪とひげの褐色の顔といった中東的特徴を持つ画像を生成する。職業に関しても，現実の統計以上に偏りを増幅し，「ソフトウェア開発者」は白人男性，「家政婦」は有色人種女性として描写される。また，イラク人は戦争，エチオピア人は飢餓といったイメージで描かれるうえに，家や自家用車を追加すると，そのステレオタイプを増幅するような画像が追加される。さらに問題なのは，プロンプトにバイアスを緩和するような「富裕な」や「大邸宅」を付け加えて「富裕なアフリカ人」「アフリカ人の大邸宅」としても，これらのステレオタイプが持続してしまうことである。

5. バイアスに対抗する戦略

生成 AI のバイアスに対処するには，プロンプト（入力）を工夫することでデータセットに内在するバイアスを打ち消すか，アルゴリズムを改良するか，データセットそのものを改良してバイアスを減らすかの主に３つの戦略がある。１つ目では「多様な」等の修飾語を付けることである程度の改善がみられるが，先に引用したように，「裕福な」や「大邸宅に住む」を付けても変わらない例もあった。各社は２つ目の，多様性を高めるようなアルゴリズムの改良も試みていることは実際に利用してみると分かる。しかし，2024年初頭，Google 社の画像生成 AI ツールが，第二次世界大戦時のドイツ兵を生成した際に，歴史的にありえないような，人種的に多様なグループとして描いたことで批判を浴びた［13］。このようにアルゴリズムの調整でどこまで解決するのかは不明であるが，３つ目のデータセットそのものはすでにインターネット上に膨大な量の蓄積があるため，多様性を増すことは簡単ではないだろう。

そもそも，どれだけ多様なら多様といえるのだろうか。政治家の場合は国民の代表であることから，性別であれば男女半々が望ましいし，他の属性についてもある程度人口統計に合わせることが望ましいというコンセンサスは得られやすいだろう。しかし，職業などはどうだろうか。社長や大統領の画像には若者も含めるべきだろうか。また，チアリーダーやアメフト選手はどのような男女比率が望ましいだろうか。こうしたケースには「正解」がないために修正もしづらいだろう。最低限言えることは生成 AI のユーザーが生成 AI の仕組みを理解し，結果に誤りやバイアスが存在しうることを十分理解して利用することが大事であることである。

［参考文献］

1) 岸本充生．(2024).「生成 AI の倫理的・法的・社会的課題（ELSI）」国立国会図書館 調査及び立法考査局 編『デジタル時代の技術と社会（令和５年度　科学技術に関する調査プロジェクト）』(pp. 165–177)．国立国会図書館．https://www.ndl.go.jp/jp/diet/publication/document/2024/index.html
2) Bengio,Y. et al.（2025）. “International AI Safety Report”（DSIT 2025/001, 2025）https://www.gov.uk/government/publications/international-ai-safety-report-2025
3) パーソナルデータ＋α研究会「プロファイリングに関する最終提言」（2022年４月22日公表）https://wp.shojihomu.co.jp/shojihomu_nbl1211
4) 総務省，経済産業省「AI 事業者ガイドライン（第1.0版）」（令和６年４月19日公表）https://www.meti.go.jp/press/2024/04/20240419004/20240419004.html
5) Regulation（EU）2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations（EC）No 300/2008,（EU）No 167/2013,（EU）No 168/2013,（EU）2018/858,（EU）2018/1139 and（EU）2019/2144 and Directives 2014/90/EU,（EU）2016/797 and（EU）2020/1828（Artificial Intelligence Act）https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
6) 岸本充生（2018）「海外の法規制及び社会動向」国立国会図書館 調査及び立法考査局 編『生体認証技術の動向と活用（平成30年度　科学技術に関する調査プロジェクト）』(pp. 51–101)．国立国会図書館．https://www.ndl.go.jp/jp/diet/publication/document/2019/index.html
7) 岸本充生（2023）「デジタルプラットフォーム上の顔写真を利用することの倫理的・法的・社会的課題（ELSI）」千葉恵美子編著『デジタル・プラットフォームとルールメイキング』日本評論社 8–42 2023年９月．
8) The Engine Room（2018）. Biometrics in the Humanitarian Sector. Commissioned by Oxfam. https://www.theengineroom.org/wp-content/uploads/2018/03/Engine-Room-Oxfam-Biometrics-Review.pdf
9) Oxfam, Oxfam’s new policy on biometrics explores safe and responsible data practice, June 24, 2021. https://views-voices.oxfam.org.uk/2021/06/oxfams-new-policy-on-biometrics-explores-safe-and-responsible-data-practice/
10) Kotek, H., Dockum, R. and Sun, D. Q.（2023）. Gender bias and stereotypes in Large Language Models. Gender bias and stereotypes in Large Language Models. In Proceedings of The ACM Collective Intelligence Conference（CI ‘23）. Association for Computing Machinery, New York, NY, USA, 12–24. https://doi.org/10.

1145/3582269.3615599
11） UNESCO, IRCAI（2024）. “Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models”. https://unesdoc.unesco.org/ark:/48223/pf0000388971
12） Bianchi, F. et al;.（2023）. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency（FAccT ‘23）. Association for Computing Machinery, New York, NY, USA, 1493–1504. https://doi.org/10.1145/3593013.3594095
13） Robertson, Adi（2024）. Google apologizes for ‘missing the mark’ after Gemini generated racially diverse Nazis. The Verge, Feb 22, 2024. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

執筆者紹介

岸本　充生

大阪大学 D3 センター（旧・データビリティフロンティア機構） 教授
大阪大学社会技術共創研究センター（ELSI センター）長

1998年京都大学大学院経済学研究科後期博士課程修了。博士（経済学）。専門はリスク学，政策評価。

通産省工業技術院資源環境技術総合研究所，独立行政法人産業技術総合研究所，東京大学公共政策大学院を経て，現在は，大阪大学 D3 センター（旧・データビリティフロンティア機構）教授。大阪大学社会技術共創研究センター（ELSI センター）長を兼任。