国立国会図書館 調査及び立法考査局

Research and Legislative Reference Bureau
National Diet Library

論題 Title	第 9 章　生成 AI の倫理的・法的・社会的課題（ELSI）
他言語論題 Title in other language	Chapter 9 Ethical, Legal, and Social Issues of Generative AI
著者 / 所属 Author(s)	岸本充生（KISHIMOTO Atsuo）／大阪大学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センター長・国立国会図書館客員調査員
書名 Title of Book	デジタル時代の技術と社会　科学技術に関する調査プロジェクト報告書（Technology and Its Social Implementation in the Digital Era）
シリーズ Series	調査資料 2023-5（Research Materials 2023-5）
編集 Editor	国立国会図書館 調査及び立法考査局
発行 Publisher	国立国会図書館
刊行日 Issue Date	2024-3-26
ページ Pages	165-177
ISBN	978-4-87582-923-2
本文の言語 Language	日本語（Japanese）
摘要 Abstract	生成系 AI が登場し、従来型 AI とは異なる倫理的・法的・社会的課題（ELSI）が顕在化した。従来型 AI の ELSI を確認した上で、生成系 AI の開発及び利用における ELSI、法規制の動向をまとめる。

第 9 章　生成 AI の倫理的・法的・社会的課題（ELSI）

大阪大学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　センター長
国立国会図書館　客員調査員　岸本　充生

目　　次

はじめに
Ⅰ　分類・推測・認識タイプの AI の ELSI
　1　学習及び推論プロセス
　2　指摘されている ELSI
　3　実施の是非の判断
Ⅱ　生成 AI の ELSI
　1　学習及び生成プロセス
　2　プロセスに沿った ELSI 概要
　3　学習用データセットの動向
Ⅲ　生成 AI のガバナンス
　1　多層的なリスクガバナンス
　2　事業者による自主的な取組
　3　各国・地域の法規制動向
おわりに

【要　旨】
2022 年半ば以降、生成タイプの AI が登場し普及し始め、それまでの分類・推測・認識タイプの AI という従来型の AI とは異なる倫理的・法的・社会的課題（ELSI）が顕在化している。これは私たちが AI に分析される対象から、AI を利用する主体に変わったことも関係している。また、生成 AI は一部の大手テック企業がリードしており、リスクガバナンスの仕組みも従来とは異なるアプローチが必要になっている。本稿では従来型の AI の ELSI を最初に振り返り、続いて生成 AI の開発と利用のプロセスに沿って指摘されている様々な ELSI を概観する。その上でガバナンスの在り方を検討する。その際には事業者による自主的な取組を取り上げた上で、欧州、英国、米国、日本における法規制も含めた動向をまとめた。

はじめに

近年、人工知能（AI）の研究開発が急速に進み、顔認識技術を始めとして、日常生活の様々な場面においても利用されるようになってきた。AI の利活用に際しては、学習（トレーニング）データの取得の場面、モデルで使われるアルゴリズム、結果の利用のされ方などの場面において、倫理的・法的・社会的課題（Ethical, Legal and Social Issues: ELSI）が生ずる可能性が指摘されている。他の様々な新規科学技術の場合と同様に、広く社会実装されると、既存の倫理規範、法規制、社会的なルールなどとギャップが生じ得るからである(1)。

2022 年には専門家でなくても利用可能な形でテキスト生成 AI や画像生成 AI が次々と発表され、私たちは AI によって分析されるだけの存在ではなく、AI を自ら利用できる存在にもなった。生成 AI の登場には、その学習（トレーニング）のための膨大な「データセット」が必要不可欠であり、デジタル化とインターネットの普及により私たちが膨大なテキストや画像をインターネット上に書き込んだり、掲載したりしてきたことがそれを可能にしたという背景がある。特に、ソーシャル・メディア・プラットフォームの登場と普及がこれを後押しした。しかし、現在ではソーシャル・メディアが生成 AI によって生み出される誤情報（ミスインフォメーション）と偽情報（ディスインフォメーション）対策に頭を悩ませることになっている。こうした背景の下で、AI ガバナンスの議論においてはにわかに「安全性（safety）」がキーワードとして浮上してきた。

本稿では、生成 AI によって生じることが懸念されている ELSI の概要をまとめる。第Ⅰ節ではこれまでの AI、すなわち、顔認識技術のような分類・推測・認識タイプの AI においてどのような ELSI が指摘されていたかについて振り返る。第Ⅱ節では生成タイプの AI について指摘されている ELSI をまとめる。第Ⅲ節では ELSI に対処するための法規制を含む取組状況について触れる。

＊本稿におけるインターネット情報の最終アクセス日は、令和 6（2024）年 2 月 3 日である。

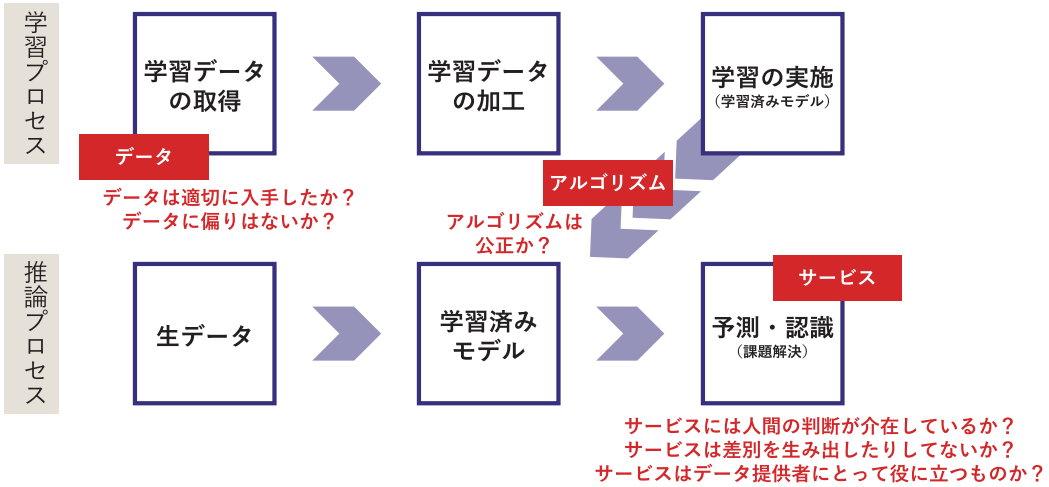
(1)　新興技術を社会実装する際に生じ得る様々な課題については、岸本充生「新興技術を社会実装するということ」国立国会図書館調査及び立法考査局編『ゲノム編集の技術と影響―科学技術に関する調査プロジェクト 2020 報告書―』（調査資料 2020-5）国立国会図書館, 2021, pp.101-121. <https://dl.ndl.go.jp/pid/11656216> を参照。生体認証技術については、国立国会図書館調査及び立法考査局編『生体認証技術の動向と活用―科学技術に関する調査プロジェクト 2018 報告書―』（調査資料 2018-6）国立国会図書館, 2019. <https://dl.ndl.go.jp/pid/11257101> を参照。

Ⅰ　分類・推測・認識タイプの AI の ELSI

1　学習及び推論プロセス

分類・推測・認識タイプの AI という、生成 AI が広く普及してからは、従来型の、あるいは、伝統的なとも形容される AI について、顔認識技術を例に振り返ってみる。私たちは日常生活で、他人の顔を見ることで、知り合いかどうか、男性か女性か、機嫌が悪そうか良さそうか、おおよその年齢、更にはどういう性格か、などの情報を推定している(2)。こうした作業はこれまでの人生における様々な経験を基にした知識を使って実施されている。そのため、人生経験が多いほどかえって固定観念に縛られてしまう可能性もある。顔認識技術の仕組みも基本的にこれと同じである。膨大な学習（トレーニング）用のデータに、目的に応じたラベル（性別、年齢、名前等）を付与して、機械学習モデルに学習させるのである。その結果、作成されたモデルに顔写真を入力すると、性別や年齢又は同一人物であるかどうかを自動的に機械が判断してくれる。人が判断する場合と同様、機械の場合も学習量が少ないケースには弱く、一定の誤った結果を出力してしまう。人の場合も機械の場合も同様に、出力結果をそのまま適用してしまうことでアンコンシャスバイアスや差別を生み出してしまう可能性を持つ。図 1 には従来型の AI の開発と利用のプロセスを示す。また、その中で ELSI が生じやすい場面として、データ取得、アルゴリズム、サービスとしての利用の 3 つの場面が挙げられる。

図 1　従来型の AI の場合の開発及び利用のプロセスと ELSI が生じる場面



（出典）筆者作成。

2　指摘されている ELSI

データが取得される場面では、個人情報保護や知的財産保護の観点から適切な手順で入手されたかという側面と、モデルの利用目的に照らしてデータセットに偏りがないかどうかという側面に注意すべきである。顔認識技術の場合は、欧米では当初、学習用データセットに含まれる顔画像が白人男性に偏っていたために、モデルのアルゴリズムに偏りが生じ、有色人種や女性において認識精度が低くなることが指摘された。先に指摘したように、モデルの出力はあく

(2)　正確には顔だけでなく、髪型や毛髪の状態、服装、歩き方、振る舞い、状況などから総合的に判断していると考えられる。

までも確率的に高い結果を出力しているにすぎず、誤りはあり得る。そのため、出力結果をそのまま個人に対する重要な意思決定に用いることには慎重であるべきである。例えば、入学や入社のための書類や面接の審査、家を購入したり事業を起こしたりするための大きな額の融資の是非の判断、罪を犯した人の刑期や出所の判断などが挙げられる。欧州（EU）の一般データ保護規則（General Data Protection Regulation: GDPR）[3]では、人間中心であるべきだとして、個人データをアルゴリズムによって自動的に処理されることに対して、異議を申し立てる権利や決定に服さない権利、知る権利などが規定されている[4]。

３　実施の是非の判断

パーソナルデータと AI を使うといろいろなことが技術的にできるようになった。そのため、「技術的にできること」と「社会的にやってもよいこと」に大きな乖離（かいり）が生じるようになった。では、当該データの利活用が、「社会的にやってもよいこと」かどうかはどうやって判断したらよいのだろうか。この線引きは状況や使い方によって変わるため、単純なガイドラインやチェックリストによって外から与えられることを期待すべきではなく、データ利活用を進める側が、データ提供者や利用者にとって受け入れられないリスクを生じさせないかしっかり検討する必要がある。実践的には、大学等で実施されている研究倫理審査のようなプロセスを経ることにより、案件ごとに、やるべきでないものからやってよいものまでのグラデーションの中に位置付ける。そのための１つの手法として、リスクアセスメントが挙げられる。情報技術のリスクアセスメントは特にプライバシー影響評価（Privacy Impact Assessment: PIA）と呼ばれる[5]。技術やサービスを導入する前に、それらのライフサイクル全体を通してありそうなリスクを洗い出した上で、発生可能性と影響の大きさの二軸で定性的に評価し、それらのリスクが受入れ可能なレベルであることを示す一連のプロセスである。

Ⅱ　生成 AI の ELSI

１　学習及び生成プロセス

生成 AI はこれまでの AI と同様、大量のデータを集めて学習（トレーニング）用データセットを作成し、それらに基づいてモデルを作成し、アプリやサービスとして提供される。生成 AI の学習用データセットのほとんどは、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データからなる。AI 開発者は、そうした既存のデータセットを無償・有償で入手し、有害情報を取り除くなどの一定のフィルタリングを行った上でモデルを作成してい

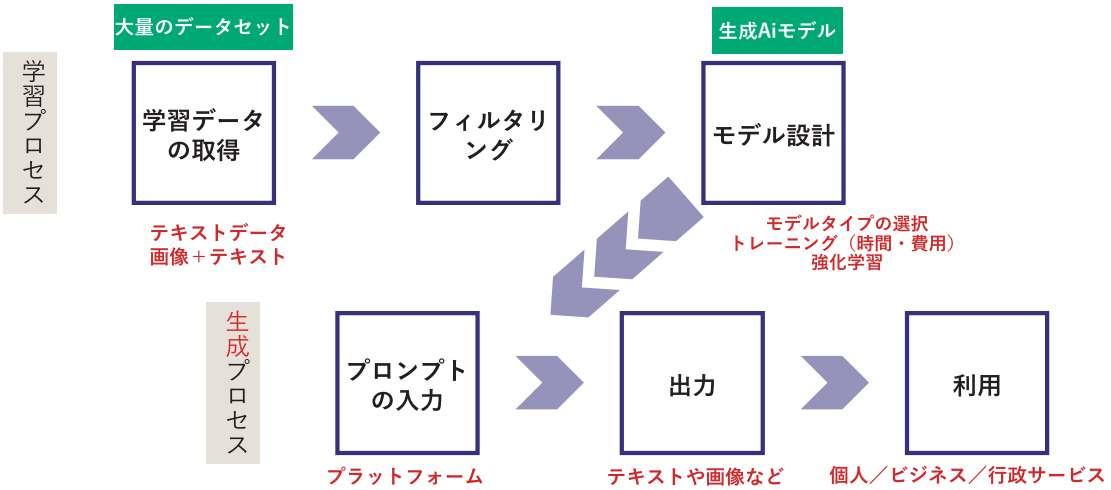
(3)　“Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation),” OJ, L119, 2016.5.4, pp.1-88.

(4)　これらはデータによるプロファイリングと呼ばれる。通常のパーソナルデータの組合せから、要配慮個人情報を推知することも技術的には可能であり、要配慮個人情報を取得しなくてもそれらを推知され得ることには注意すべきである。プロファイリングに対する規制は日本国内にはまだないが、自主的取組のためのチェックリストが専門家により提案されている。パーソナルデータ＋α研究会による「「プロファイリングに関する最終提言」の公表」商事法務ポータルウェブサイト <https://wp.shojihomu.co.jp/shojihomu_nbl1211> を参照。ここではプロファイリングを「パーソナルデータとアルゴリズムを用いて、特定個人の趣味嗜好、能力、信用力、知性、振舞いなどを分析又は予測すること」と定義している。

(5)　岸本充生「デジタル化に伴う ELSI とリスクコミュニケーション」奈良由美子編著『リスクコミュニケーションの探究』放送大学教育振興会, 2023, pp.254-273.

る。利用者がテキストなどをプロンプトとして入力することでテキスト、画像、プログラムコードなどが自動的に生成される。図 2 には、AI 開発者又は提供者による学習プロセスと AI 利用者による生成プロセスという単純化した 2 つのプロセスを記載した。実際には両者の間に、独自データを加えてファインチューニングを施して独自モデルを作ってサービスを行う AI サービス事業者が加わっている場合も多い。その場合、AI 開発者が開発するモデルは汎用目的の AI モデルであるため基盤モデルと呼ばれることもある。Ⅱ 2 では生成 AI の学習及び生成プロセスにおいて指摘されている ELSI をプロセスに沿ってまとめる。

図２　生成 AI の場合の開発及び利用のプロセス



（出典）筆者作成。

２　プロセスに沿った ELSI 概要

(1) 学習データの取得

著作権：学習用のデータセットについては、「情報解析の用に供する場合」として、国内では著作権法（昭和 45 年法律第 48 号）第 30 条の 4 の権利制限規定が適用されるとされている。しかし、「著作権者の利益を不当に害する場合」を除くとするただし書きがあり、これは「著作権者の著作物の利用市場と衝突するか」あるいは「将来における著作物の潜在的市場を阻害するか」という観点から判断されるとされているものの[6]、生成 AI の場合に具体的にどこまで適用されるかについては議論が続いている。米国では、機械学習に利用される場合には「フェアユース」概念が適用されるとされている。また、大規模言語モデルでよく利用されているデータセットに海賊版の書籍が多く含まれているケースも指摘されている[7]。しかし、通常は、海賊版であるか否かは著作権者でなければ判断が困難であり、意図的に利用している場合を除き、AI 開発事業者の責任を問うのは難しいとも指摘されている[8]。

個人情報：国内法では要配慮個人情報でなければ利用目的の通知で十分であり、同意は必ず

(6)　文化庁著作権課「デジタル化・ネットワーク化の進展に対応した柔軟な権利制限規定に関する基本的な考え方について」2019.10.24, p.24. <https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/pdf/r1406693_17.pdf>

(7)　Alex Reisner, “These 183,000 Books are fueling the biggest fight in publishing and tech,” *The Atlantic*, September 25, 2023. <https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/>

(8)　文化庁文化審議会著作権分科会法制度小委員会「AI と著作権に関する考え方について（素案）（令和 6 年 1 月 23 日時点版）」

しも必要ない。そのため個人情報保護委員会は2023年6月、ChatGPTを提供するOpenAI社に対して機械学習のために収集する情報に要配慮個人情報が含まれないように注意喚起を行った[9]。他方、欧州（EU）のGDPR第6条ではデータ処理のための法的根拠が6つ挙げられており、「(a) 同意」を個別にとることは現実的に困難であるために、OpenAI社などは「(f) 正当な利益（legitimate interests）」を法的根拠として挙げている[10]。しかし、この主張についてはEU当局及び加盟国当局が精査中であり、すんなりと認められるかどうかは不明である。また、顔情報を含む生体データは日本では通常の個人情報に分類されているが、GDPRでは第9条における「特別カテゴリーの個人データ」に含まれていることにも注意が必要である。

データの偏り：データソースのほとんどがインターネット上に存在しているものであることは、データ化されていなかったり、公開されていなかったりするものはデータセットに含まれていないことを意味する。つまり、学習データが、データ化されて公開されているものに偏っているのである。また、マジョリティによるコンテンツが多くを占めることになるのも自然である。テキストであれば英語など、画像であれば西洋のアートやデザインが多くを占めることになる。Ⅱ1（5）で後述するが、インターネット上の文章や画像には現実社会に存在するバイアスがそのまま反映されていることにも注意すべきである。

透明性：著作権や個人情報の保護の観点からも、データの偏りを確認するためにも、学習用データセットの詳細を公開すべきだという議論がある。つまり、どのようなデータセットで学習やフィルタリングを行っているのかが分からないと、出力結果のバイアスが予想できず、出力結果を利用する際に何に気を付ければよいのかが分からないという批判である。しかし、例えばOpenAI社はGPT-4をリリースした際に、「競争の状況及び安全性への影響」を所与とすると、学習用データセットを含むモデルの詳細を公開しないこととしたと説明している[11]。つまりデータセットや、更にフィルタリングやファインチューニングといったモデル作成はそれ自体が付加価値であり、企業秘密情報であるとしている。また、詳細を公開することで悪用しようとする主体にヒントを与えてしまう可能性も指摘されている。透明性の重要さについては異論がなくても「どの程度透明ならば十分透明か」という具体的な議論が今後必要になる。

ただ乗り：たとえ著作権法における権利制限規定が適用される（すなわち、ただし書きが適用されない）としても、生成AI開発事業者は他人の作成した文章や画像にただ乗りして商業的に利用しているという批判は十分にあり得るし、アーティストらへの補償や利益還元を望む態度も理解できる。

(2) フィルタリング

労働者搾取：生成AIの開発事業者がモデルの安全性に力を入れていることは度々強調されている。そのために、有害情報にラベルを貼ったり、有害情報か否かを見分けたり、有害情報を取り除いたりする作業が必須であり、これらは機械による作業が十分機能するまでは、ある

(9)「生成AIサービスの利用に関する注意喚起等について」2023.6.2. 個人情報保護委員会ウェブサイト <https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/>

(10) 実際、イタリアのデータ保護当局（GPDP）は2023年3月31日、GDPRに基づき、OpenAI社に対してイタリア人ユーザーのデータ処理を直ちに停止するよう命令を出すとともに、調査を開始したことを発表した。その理由の1つに、トレーニングデータを大量に収集・処理することを裏付ける法的根拠がないように見えることを挙げた。これに対して、4月にOpenAI社は対策を回答し、イタリア国内のサービスも暫定的に再開された。その際に、「(f) 正当な利益」を法的根拠として挙げた。

(11) OpenAI, “GPT-4 Technical Report,” 27 Mar 2023. <https://cdn.openai.com/papers/gpt-4.pdf>

いは、機械が自動的に実施したとしても最終的には人による作業に依存している。こういった作業の多くは途上国の労働者に低賃金で外注されており、彼らの多くがこうした作業を通して心的外傷を負っていることが報道された[12]。このような労働は「ゴースト・ワーク」と呼ばれることもある[13]。

(3) モデル設計

エネルギー・資源：生成 AI モデルの作成や利用には大きな計算資源（コンピューティング・パワー）が必要であり、そのため膨大な電力を消費し、二酸化炭素排出を含む環境影響が非常に大きいことが指摘されている[14]。また、データセンターの冷却に大量の淡水資源が必要であることや、ハードウェアにはレアアース類を含む金属や鉱物が大量に必要となることも指摘されている。

独占・寡占：生成 AI モデルの作成には膨大な計算資源が必要であることが参入障壁となり、自由な市場競争が妨げられるおそれがある。また、著作権や個人情報保護などの観点から大規模な学習用データセットの自由な利用が今後難しくなるならば、すでに汎用目的の AI モデルを作成した先行事業者に対して、新規参入が難しくなる可能性がある。

ブラックボックス：通常の AI と同様、モデルの透明性が課題となり得る。入力と出力の関係、つまりアルゴリズムが不明なモデルにおいては、結果の説明可能性や解釈可能性が十分でないおそれがあり、責任のある利用ができないことになる。

(4) プロンプトの入力

著作権：プロンプトの入力に関しても日本の著作権法第 30 条の 4 の権利制限規定は適用される。しかし、著作権を侵害するような出力を意図した場合には、これは適用されない。

情報漏洩（ろうえい）：プロンプトに個人情報や企業秘密情報を入力した場合に、利用規約によっては、学習目的で利用される可能性があり、その場合に出力情報に含まれることで情報漏洩につながる可能性が指摘されている。

脱獄プロンプト：チャットボットなどの対話型 AI の脆弱（ぜいじゃく）性を狙ったプロンプトを入力し、倫理的なセーフガードを回避して、有害なコンテンツや個人情報を引き出したりする攻撃が確認されている。

(5) モデルの出力

誤情報（ミスインフォメーション）：大規模言語モデルによって生成される誤情報は、「幻覚（hallucination）」と呼ばれることがある。これは確率的にありそうな単語列を生成するという言語モデルの性質上避けて通れないリスクであり、むしろ現実よりも「ありそうな」文章が生成される。実際に、米国において ChatGPT を「スーパーサーチエンジン」と勘違いした弁護

---

[12] TIME 誌が 2023 年 1 月 18 日、OpenAI 社が、学習データから有害なコンテンツを取り除くための作業を時給 2 ドル以下でケニア人労働者に外注していたことを明らかにした。Billy Perrigo, “Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic,” *Time*, JANUARY 18, 2023. <https://time.com/6247678/openai-chatgpt-kenya-workers/>

[13] メアリー・Ｌ . グレイ, シッダールタ・スリ（柴田裕之訳）, 成田悠輔監修『ゴースト・ワーク―グローバルな新下層階級をシリコンバレーが生み出すのをどう食い止めるか―』晶文社, 2023.（原書名 : Mary L. Gray and Suri Siddharth, *Ghost Work*, New York: Harper Business, 2019.）

[14] Niklas Sundberg, “Tackling AI’s Climate Change Problem,” *MIT Sloan Management Review*, Vol.65 No.2, pp.38-41.

士が、ChatGPT が出力した存在しない判例を引用した文章を裁判所に提出してしまった例が有名である[15]。こうした事態を避けるためにも生成 AI の仕組みをある程度理解してから利用することが必須である。

バイアス：先にも書いたように、学習用データセットがインターネット上の文章あるいは画像である以上、出力が、現実に存在するステレオタイプやバイアスを再現してしまうことは容易に想像できる。画像生成 AI において、大統領や CEO、社長など、組織の長を表す言葉を入力するとほとんど白人中高年男性の画像が出力される。Bloomberg の記事によると、現実のバイアスが単に「再現」されるのではなく、「拡大」することが指摘されている[16]。米国の裁判官の 34% が女性であるにもかかわらず、「裁判官」というキーワードで生成された画像のうち、女性だと認識できた画像はわずか 3% であったという。また、人物描写だけでなく、建物や自動車、街並みといった画像の背景にある要素にも様々なバイアスが反映していることも指摘されている[17]。

文化的周縁化：データ化された画像を学習用データセットとしている以上、アートの分野においては白人による西洋アートが支配的であり、AI が生成するアートは当然西洋的なものになりがちである。これは音楽や文学にも当てはまる。そのため、周縁化された芸術は画像生成 AI からは排除されることになる。もちろん、データ化されていないために AI 生成モデルから「保護される」という見方もできるが、「存在しないもの」とされてしまうリスクも抱えることになることにも注意が必要である。

(6) 出力の利用

著作権：日本国内では生成・利用段階における著作権の考え方は AI の有無に関係なく、類似性と依拠性の両者が認められる場合に著作権侵害となる。生成 AI において特に問題となるのは依拠性の判断であり、特に「AI 利用者が既存の著作物を認識していなかったが、AI 学習用データに当該著作物が含まれる場合」の判断である。この場合は、客観的に当該著作物へのアクセスがあったと認められることから依拠性があったと推認されることになるとされている[18]。また、AI 生成物が著作権法による保護を受けるのかどうかについては、指示・入力の分量・内容、生成の試行回数、複数の生成物からの選択などを総合的に考慮して判断されるとされている。

擬人化：孤独対策やメンタルヘルスケアなどの目的でチャットボットを擬人化する試みも行われている一方で、擬人化することで「幻覚」や詐欺を信じやすくなってしまう可能性がある。つまり、人は擬人化されることで感情的に操作されやすくなり、自律性を損ない、より脆弱な存在になってしまう。そうした懸念から、擬人化を防ぐためにチャットボットにおいて絵文字の使用をやめるべきだという提案もある[19]。チャットボットを安易に擬人化せずに、ユーザーに

---

⑮　Benjamin Weiser and Nate Schweber, “The ChatGPT Lawyer Explains Himself,” *New York Times,* June 8, 2023. <https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html>

⑯　Leonardo Nicoletti and Dina Bass, “Humans are biased. Generative AI is even worse,” *Bloomberg*, June 8, 2023. <https://www.bloomberg.com/graphics/2023-generative-ai-bias/>

⑰　Federico Bianchi et al., “Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale,” *FAccT ’23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp.1493-1504. <https://doi.org/10.1145/3593013.3594095>

⑱　文化庁文化審議会著作権分科会法制度小委員会　前掲注 (8)

⑲　Carissa Véliz, “Chatbots shouldn’t use emojis,” *Nature*, Vol.615, 16 March 2023, p.375. <https://doi.org/10.1038/d41586-023-00758-y>

機械と対話していることを常に意識させる方が私たちの自律性を守るという点で倫理的であろう。

偽情報（ディスインフォメーション）：悪意ある行為者が詐欺やプロパガンダを行うために必要なコストを大きく下げてしまうことが指摘されている。文章と画像を組み合わせた「フェイクニュース」、偽の声を使う「ボイスクローン」、偽の動画を作成する「ディープフェイク」などが問題になっている。ボイスクローン技術を使って誘拐を偽装し身代金をだまし取ろうとする詐欺行為などがすでに発生している。政治的に利用されれば民主主義の基盤を揺るがしかねないことで、欧米で大きな選挙のある 2024 年は特に警戒が高まっている。生成 AI サービス各社の利用規約や利用方針において、こうした利用形態は禁止事項として挙げられているものの、それだけでどこまで抑止力となるかは疑問である。

差別：ステレオタイプやバイアスが拡大再生産された出力結果を利用することで、既存のステレオタイプを固定化させ、差別を拡大再生産する可能性がある。

デュアルユース：軍事利用も懸念されている。米国防省は 2023 年 8 月、生成 AI タスクフォースを設立し、生成 AI のユースケースを検討している[20]。

3　学習用データセットの動向

ChatGPT を始めとする多くのテキスト生成 AI では、非営利団体コモンクロール（Common Crawl）がウェブスクレイピングにより作成したデータセット（特に C4 データセットと呼ばれるものが有名である。）や書籍データ、ウィキペディアなどを学習（トレーニング）用データセットとして利用している。C4 データセットは一定のフィルタリングが掛けられているものの、問題あるコンテンツが含まれていることが度々指摘されている[21]。Meta 社が開発した大規模言語モデルである LLaMA などの学習（トレーニング）データとしても利用された書籍データセットである Books3 は非営利の研究グループである EleutherAI が作成したものである。20 万冊近くの書籍が含まれていた。しかし、デンマークの反著作権侵害団体である Rights Alliance が 2023 年 8 月、これらの書籍のほとんどが海賊版であることから、Books3 データセットをホストしていたリポジトリに対して削除要請を通知し、すぐに削除された[22]。しかし、一度流通したデータセットをインターネット上から完全に削除することは困難であり、様々なサイトから利用可能なままであることが指摘されている。

Stable Diffusion を始めとする多くの画像生成 AI では、LAION と呼ばれるドイツの非営利団体が作成した約 58 億 5000 万個のカラー画像とテキストのペアからなる研究用データセット「LAION-5B」が用いられている。これは 100% 非営利であり、教育研究用に無償で提供されていた。もともと「キュレーションされていないというデータセットの性質上、収集されたリンクは、人間の閲覧者にとって強い不快感や不穏な気分を与えるコンテンツにつながる可能性があることに留意してください。したがって、デモ・リンクの使用は慎重かつ自己責任でお願いします。」という注意書き[23]があったが、スタンフォード大学の研究プロジェクトが 2023 年 12

---

[20] “DOD Announces Establishment of Generative AI Task Force,” Aug. 10, 2023. U.S. Department of Defense Website <https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force/>

[21] Kevin Schaul et al., “Inside the secret list of websites that make AI like ChatGPT sound smart,” *Washington Post*, April 19, 2023. <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/>

[22] Ernesto Van der Sar, “Anti-Piracy Group Takes Prominent AI Training Dataset “Books3” Offline,” *TorrentFreak*, August 16, 2023. <https://torrentfreak.com/anti-piracy-group-takes-prominent-ai-training-dataset-books3-offline-230816/>

[23] Romain Beaumont, “LAION-5B: A new era of open large-scale multi-modal datasets,” 31 Mar, 2022. LAION Website <https://laion.ai/blog/laion-5b/>

月、1,000件を超える児童性的虐待コンテンツ（Child Sexual Abuse Material: CSAM）が含まれていることを指摘したため[24]、LAIONはすぐにデータセット全体を一時的に削除し、安全を確認してから再公開することを発表した[25]。しかし、LAION-5Bデータセットはすでに広く普及しているために利用を止めることは事実上不可能である。

Ⅲ　生成AIのガバナンス

1　多層的なリスクガバナンス

2023年5月に開催されたG7広島サミットの結果を踏まえて、生成AIを含む「高度なAIシステム」に関する国際的なルール作りを行うために「広島AIプロセス」が立ち上げられた。OpenAI社を始めとする生成AI事業者自身や後述する英国政府は、これらを「フロンティアAI」とも呼んでいる。生成AIが社会に普及するにつれて、Ⅱで取り上げたような様々なリスクが指摘されたり顕在化したりしたため、2023年はそれらの「安全性」に注目が集まった。生成AI開発事業者による自主的取組と、国や地域（EU）レベルの法規制の提案、国際的な原則やガイドラインの作成など、多層的な取組が同時並行で進んでいる。

生成AIシステムのガバナンスのモデルとして、OpenAI社は自社ブログにおいて将来的には国際原子力機関（IAEA）のような監査を実施できる国際機関を設けることを提案した[26]。エイダラブレス（Ada Loveless）研究所からは、米国の食品医薬品局（FDA）が生命科学（特にクラスⅢの医療ソフトウェア）に対して課している規律が適用できるかという論考も公表された[27]。

2　事業者による自主的な取組

米国の生成AIの主要企業であるAmazon社、Anthropic社、Google社、Inflection社、Meta社、Microsoft社、OpenAI社の7社は2023年7月、米国政府からホワイトハウスに招集され、AI技術がもたらすリスクの管理について、AIの未来にとって基本となる安全・セキュリティ・信頼という3つの原則を強調する自主的なコミットメントを約束した[28]。その直後に、Anthropic社、Google社、Microsoft社、OpenAI社の4社はフロンティアAIモデルの安全で責任ある開発を確保するための新しい業界団体である「フロンティア・モデル・フォーラム（Frontier Model Forum）」を発足させた[29]。OpenAI社はその前に、公共の安全に深刻なリスク

---

(24)　David Thiel, “Investigation Finds AI Image Generation Models Trained on Child Abuse,” December 20, 2023. The Cyber Policy Center Website <https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse>

(25)　LAION.ai, “Safety review for LAION 5B,” 19 Dec, 2023. <https://laion.ai/notes/laion-maintanence/> なお、再公開時期は未定である。

(26)　Sam Altman et al., “Governance of superintelligence,” May 22, 2023. OpenAI Website <https://openai.com/blog/governance-of-superintelligence>

(27)　Merlin Stein and Connor Dunlop, “Safe before sale: Learnings from the FDA’s model of life sciences oversight for foundation models. Discussion Paper,” 14 December 2023. Ada Loveless Institute Website <https://www.adalovelaceinstitute.org/report/safe-before-sale/>

(28)　“FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI,” July 21, 2023. The White House Website <https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/> 9月には第2ラウンドとして、Adobe社、IBM社など更に8社と新たに自主的コミットメントを約束したことを発表した。

(29)　Frontier Model Forum Website <https://www.frontiermodelforum.org/>

をもたらす能力を保有する可能性のある基盤モデルを「フロンティア AI」と呼び、自主規制だけでは人々を十分に保護できる可能性は低いために、政府による介入が必要になると主張していた[30]。OpenAI 社は 9 月には、モデルの安全性を向上させるため、「OpenAI レッドチーミング（Red Teaming）ネットワーク」に参加する様々な分野の専門家の募集を開始した[31]。専門家として必要なドメインは、認知科学、生物学、コンピューターサイエンス、政治科学など 26 分野が挙げられた。Anthropic 社も 9 月、「責任あるスケーリング・ポリシー（Responsible Scaling Policy: RSP）」の Version 1.0 を発表した[32]。そこでは、AI モデルが高度になればなるほどリスクも増大することが想定されることから、「バイオセーフティレベル（Biosafety Level: BSL）」基準を参考にした、破局的リスクに対処するための 4 段階の「AI 安全レベル（AI Safety Level: ASL）」の仕組みが提案された。OpenAI 社は 10 月、フロンティア AI モデルがもたらすかもしれない深刻なリスクを軽減するために「準備（Preparedness）チーム」を発足させた。準備チームの任務には、「リスク情報に基づく開発方針（Risk-Informed Development Policy: RDP）」の策定と維持も含まれている。

他方で代替的なビジネスモデルを模索する動きもある。例えば、Adobe 社は 2023 年 3 月、Adobe ストックの画像、オープンライセンスコンテンツ、著作権が失効したパブリックドメインコンテンツを利用した「商用利用にも安全なコンテンツを生成するように設計」された画像生成 AI モデルである Adobe Firefly を発表した。さらに、企業顧客が万が一、著作権侵害で訴えられた場合には、Adobe 社が法的訴訟を引き受け、請求に対して幾らかの金銭的な補償を提供する予定であるとした。

3　各国・地域の法規制動向

(1) 欧州

2021 年 4 月に欧州委員会から AI 規則案が提案されたが、2022 年に生成 AI が急速に普及したことを踏まえて、これらを加味した形で 2023 年 6 月には欧州議会による修正案が採択された。その後、12 月には議会・理事会・委員会による三者協議（トリローグ）における合意に達し、2024 年 4 月にも最終版が公表される予定である[33]。EU 官報に掲載されてから 20 日後に発効し、禁止事項に関する措置が最も早くて 6 か月後から適用されるとされている。

案では、リスクに基づくアプローチが採用され、リスクの大きさに応じた規制内容となっている。許容できないリスクを持つとされるカテゴリーの AI システムの実践は禁止された。サブリミナル技術を用いたり、意図的に操作したり騙（だま）したりするもの、年齢や障害といった脆弱性を利用するもの、生体データからセンシティブな特性を推論する分類システム、ある種の社会的スコアリングシステム、法執行目的で公共空間において行われるリアルタイムの遠

---

[30] “Frontier AI regulation: Managing emerging risks to public safety,” July 6, 2023. OpenAI Website <https://openai.com/research/frontier-ai-regulation>

[31] “OpenAI Red Teaming Network,” September 19, 2023. OpenAI Website <https://openai.com/blog/red-teaming-network> ここでのレッドチーム（Red Team）とは、AI システムについての様々な範囲のリスクアセスメントを包含する用語として使われている。

[32] “Anthropic’s Responsible Scaling Policy,” September 19, 2023. Anthropic Website <https://www.anthropic.com/news/anthropics-responsible-scaling-policy>

[33] 以下で述べる AI 規則案の内容は EU 加盟国が 2024 年 2 月 2 日に合意した際のテキストに基づいている。今後微修正がなされる可能性がある。Council of the European Union, “Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts,” 26 January 2024. <https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf>

隔生体識別システム、プロファイリングによって犯罪可能性を予測するシステム、などが挙げられた。次に、健康、安全、基本的権利に対して重大なリスクを課すことが懸念される「高リスクAIシステム」に分類されるAIシステムには適合性評価手続が義務付けられた。社会実装前には、「基本的権利影響評価」と呼ばれるリスクアセスメントが実施されなければならない。高リスクAIシステムは具体的には附属書Ⅲに8分野が挙げられており、健康、安全、基本的権利を害する重大なリスクを課さない場合のみ高リスクでないとみなされることになっている。そのほか、AIチャットボットのようなアプリには透明性などの要件が課された。

生成AIについては、汎用目的AIモデルのうち、高いインパクト能力を持っている場合は「システミック・リスク[34]を伴う汎用目的AIモデル」という分類が新たに設けられた。高いインパクト能力とは、FLOPsで計測される計算処理能力が10の25乗を上回ることと定義された。汎用目的AIモデルに課される義務の中には、「（今後設置される）AI Officeが提供するテンプレートに従って、学習に使用したコンテンツに関する十分詳細な要約を作成し、一般に公開すること」という項目がある。

（2）英国

英国は、伝統的な規制アプローチを推進する欧州（EU）とは対照的に、AIを含む新規科学技術に関して、自主的取組を中心にイノベーション促進的なアプローチを推進してきた。しかし生成AIの登場を受けて、2023年6月、リシ・スナク（Rishi Sunak）首相は英国においてAIの安全性に関する「初のグローバルサミット」の開催を宣言し、AIのリスクガバナンスの主導権を取る方向へ軌道修正した。科学・イノベーション・技術省（DSIT）は6月に「基盤モデルタスクフォース（Foundation Model Taskforce）」を設置したが、9月にはその名称が「フロンティアAIタスクフォース（Frontier AI Taskforce）」に変更された。11月1～2日にはロンドンのブレッチリー・パーク（Bletchley Park）においてAI安全サミットが開催された。会議での主な対象を「フロンティアAI」と「危険な能力を持つ狭いAI」とし、フロンティアAIは、「多種多様なタスクを実行でき、今日の最先端モデルに存在する能力と同等か、それ以上の能力を持つ、非常に能力の高い汎用目的AIモデル」と定義された[35]。

（3）米国

米国ではこれまで法的拘束力のない「AI権利章典のためのブループリント」[36]と、国立標準技術研究所（NIST）による自主的取組のためのガイドラインである「AIリスクマネジメント枠組み」[37]を中心に議論が進められてきたが、生成AIの登場により2023年からは大手テック企業による自主的取組を約束させる共同規制的なアプローチを推進し始めた。さらに、ジョー・バイデン（Joe Biden）大統領は10月30日、「安全で、セキュアで、信頼できるAIに関する大

⑶④　システミック・リスクとは、EU域内市場全体に、公衆衛生、安全、治安、基本的権利又は社会全体に対して重大な影響を及ぼす、汎用目的AIモデルの高インパクト能力に特有のリスクのことを指している。

⑶⑤　UK Government, “Introduction to the AI Safety Summit,” [September 25, 2023], p.[2]. <https://assets.publishing.service.gov.uk/media/65255655244f8e00138e7362/introduction_to_the_ai_safety_summit.pdf>

⑶⑥　“Blueprint for an AI Bill of Rights,” October 2022. The White House Website <https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>

⑶⑦　“Artificial Intelligence Risk Management Framework (AI RMF 1.0),” January 2023. NIST Website <https://doi.org/10.6028/NIST.AI.100-1>

統領令（Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence）」を公布した[38]。AI の安全性とセキュリティのために、最も強力な AI システムの開発者に対して安全性試験の結果やその他の重要な情報を米国政府と共有することを義務付けたほか、NIST に AI システムの安全性、セキュリティ、信頼性を確保するための基準、ツール、試験を開発させるなど、連邦政府省庁に対して様々な新基準を策定することを指示した。

（4）日本

日本政府は 2023 年 4 月に「AI 戦略チーム（関係省庁連携）」を、5 月には有識者からなる「AI 戦略会議」を立ち上げた[39]。2023 年 5 月に日本の広島で G7 サミットが開催され、首脳声明には「責任ある AI」の推進に向けて、「広島 AI プロセス」の創設が盛り込まれた[40]。実際に実務者レベルの作業部会が 5 月末に開始され、10 月 30 日に「広島 AI プロセスに関する G7 首脳声明」に加えて、「高度な AI システムを開発する組織向けの広島プロセス国際指針」及び「高度な AI システムを開発する組織向けの広島プロセス国際行動規範」が公表された[41]。

国内向けには 10 月末に経済産業省と総務省が、AI 事業者ガイドライン検討会を立ち上げ、2024 年 1 月に法的拘束力のない「AI 事業者ガイドライン案」が公表され、パブリックコメントにかけられた[42]。

おわりに

AI のリスクガバナンスに関する議論は、生成 AI の登場によって大きく様相を変え、「安全性」が前面に出てきた。安全性の議論は時に、長期的に人類存続へのリスクを課すかどうかというアジェンダを浮上させるようになった。英国が主催した AI 安全サミットも当初はそうした議題が前面に出ていた印象であった。これに対して、人類存続リスクは AI ハイプ（技術楽観主義とも密接に関係している。）の裏返しであり、現在すでに指摘されたり顕在化したりしている差別やバイアスから目を逸らせる効果があり、これらの課題こそ優先的に取り組むべきであるという批判も多い。結果として、AI 安全サミットでは後者にも十分に時間を割いたものになった。

（きしもと　あつお）

---

[38] “FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence,” October 30, 2023. The White House Website <https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/> 後に、Executive Order 14110 という番号が付けられた。

[39] 「AI 戦略会議」内閣府ウェブサイト <https://www8.cao.go.jp/cstp/ai/ai_senryaku/ai_senryaku.html>

[40] 「（仮訳）G7 広島首脳コミュニケ（2023 年 5 月 20 日）」pp.27-29. G7 広島サミットウェブサイト <https://www.g7hiroshima.go.jp/documents/pdf/Leaders_Communique_01_jp.pdf?v20231006>

[41] 「広島 AI プロセスに関する G7 首脳声明」2023.10.30. 外務省ウェブサイト <https://www.mofa.go.jp/mofaj/ecm/ec/page5_000483.html>

[42] 「AI 事業者ガイドライン検討会」経済産業省ウェブサイト <https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/index.html>