大阪大学 社会技術共創研究センター
Research Center on Ethical, Legal and Social Issues

ELSI NOTE No.26

2023 年 4 月 10 日

生成AI（Generative AI）の倫理的・法的・社会的課題（ELSI）論点の概観

2023年3月版

Authors

カテライ アメリア	大阪大学　社会技術共創研究センター（ELSIセンター）	特任助教（2023年04月現在）
井出　和希	大阪大学　感染症総合教育研究拠点/ELSIセンター	特任准教授（2023年04月現在）
岸本　充生	大阪大学　社会技術共創研究センター（ELSIセンター）	センター長（2023年04月現在）

Acknowledgements

本 NOTE プロジェクトは、2023 年 1 月 6 日の岸本からの「生成 AI の ELSI について気軽にしゃべる会を開催したいと思います」という呼び掛けに応じた社会技術共創研究センター（ELSI センター）メンバー有志と、1 月 16 日に集まって意見交換したことから始まりました。その後も、ELSI センターメンバーからは断続的にコメントや情報提供を受け、また、議論を交わしました。関係者に感謝します。

また、脚注に記載されている URL の多くは 2023 年 3 月 31 日段階でアクセスできたものであり、その後にアクセスができなくなる可能性がある。

大阪大学

目次

はじめに

生成 AI は、通常の機械学習と同様、過去の膨大な素材を学習することにより AI モデルが作成される。テキスト生成モデルである ChatGPT などは、インターネット上の膨大な量のテキストにより訓練された。画像生成モデルである Stable Diffusion などは、インターネット上の膨大な量の画像により訓練された。すなわち、人間が過去に創造したコンテンツに基づいて新しいコンテンツを生成していることになる。そのため、誤った情報、バイアス、人種やジェンダーによるステレオタイプなどを再生産してしまう可能性がある。生成 AI の場合はさらには、データ化されていないものは存在しないものとされることになることにも注意すべきである。すなわち、データ化されていないテキスト、画像、アートなどが除外されていることを意識し、そのことの含意を検討する必要がある。つまり、これまで AI の ELSI として指摘されてきたもの、すなわち、1）学習（教師）データの収集や加工は適正に行われたか、2）アルゴリズムは正確で公正なものであるか、3)製品やサービスのアウトプットや使い方は適正か、に加え、生成 AI 特有の ELSI を検討する必要がある。また、汎用目的型 AI としての性格を持ち合わせる生成 AI は、その利活用が広がることによる社会の各セクターへの短期的から中長期的な影響についても検討する必要があるだろう。

本 ELSI NOTE は、研究開発の進展やそれらへの社会の対応が急速に変化する生成 AI の分野[1]の 2023 年 3 月までの ELSI 動向を切り取ったノートである。1 週間後、1 か月後、半年後、1 年後にはずいぶん違った状況になっている可能性が高いことに注意すべきである。

1. 画像生成 AI（Text-to-image AI）

1.1. 動向

■ カリフォルニア州サンフランシスコに本社がある OpenAI 社は 2021 年 1 月に、テキスト（プロンプト）を画像に変換してくれる AI モデルである DALL-E を発表した。しかし一般ユーザー向けに公開することには、嫌がらせやプロパガンダなどの悪用を招きかねないとして慎重

1 本 NOTE は多様な生成 AI のうち、Text-to-Image と Text-to-Text を主な対象としている。生成 AI の全体像については、次のプレプリント論文の分類などを参照。Gozalo-Brizuela R, Garrido-Merchan EC ChatGPT is not all you need A State of the Art Review of large Generative AI models arXiv 2023 doi: https://doi org/10 48550/arXiv 2301 04655

であった[2]。そのため、学習データセットから「暴力的な画像」と「性的な画像」を削除し、同様のプロンプトに基づく画像生成を拒否するなど、悪用を防ぐため様々な施策を講じた[3]。また、人種やジェンダー等のバイアスを緩和するための手法を開発した。こうした対策を経て、OpenAI 社は 2022 年 9 月に DALL-E 2 を発表し、誰でも利用できるようにした。

■ 米国の Midjourney 社は同名の、テキストから画像を生成する AI プログラムを開発し、2022 年 7 月にオープンベータ版を公表した。David Holz 氏が創業者であり CEO である。オープン・バイ・デフォルトのコミュニティと称している。ユーザーが Discord からテキストで指示を送ると画像を生成するサービスである。対象年齢は 13 歳以上とされている。インターネットに公開された画像を教師データとして利用しているが、多くは著作権で保護されているものであり、同意や許可を得ているわけではない。行動規範とコミュニティガイドラインが公開されており、行動規範は次の 3 点からなる。

ー 嫌な奴にならないでください。

ー 私たちのツールを使って、煽ったり、動揺させたり、ドラマを引き起こす[4]ような画像を作らないでください。これには、グロやアダルトコンテンツも含まれます。

ー 他の人やチームに対して敬意を払ってください。

■ 2022 年 8 月 22 日に Stable Diffusion は、Stability AI 社の資金提供により、オープンソースで、フィルターされない画像生成を提供する画像生成 AI プログラムとして公開された[5]。DALL-E 2 と違って、コンテント・モデレーションがなされていないため、暴力的な画像や性的な画像、肖像権や著作権に触れそうな画像なども生成できてしまう。Stable Diffusion で生成された画像のための検索エンジンである Lexica[6]で検索できる。CEO の Mostaque 氏は「最終的には、この技術をどのように運用するか、倫理的、道徳的、合法的であるかどうかについては、人々の責任です。」とコメントしている[7]。

---

[2] https://www theverge com/2022/9/28/23376328/ai-art-image-generator-dall-e-access-waitlist-scrapped

[3] https://openai com/blog/dall-e-2-pre-training-mitigations

[4] 「ドラマを引き起こす」というのは政治的な波風を起こすことなどを指している。
https://www washingtonpost com/technology/2023/03/30/midjourney-ai-image-generation-rules/

[5] https://stability ai/blog/stable-diffusion-public-release

[6] https://lexica art/

[7] https://www theverge com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data

■ 2023 年 3 月、Adobe 社は、画像生成 AI「Adobe Firefly」のベータ版を発表した[8]。アカウンタビリティ、責任、透明性からなる「AI 倫理原則」に基づいているとされる[9]。最初のモデルは、Adobe Stock 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でトレーニングされた[10]。また、モデルのアウトプットについても、社内の AI 倫理チームが主導して、エンジニアリングチームとともに問題解決を行う体制を作っているという。

■ 2023 年 3 月、Microsoft 社は「Bing Image Creator」を発表した[11]。Bing Image Creator は、OpenAI 社のパートナーによる DALL·E モデルの高度なバージョンを搭載しており、見たい絵を自分の言葉で説明するだけで画像を作成することができる。これにより、チャット内で、文章と映像の両方のコンテンツを一度に生成できるようになった。なお、Microsoft 社では、チームが責任を持って AI システムを開発・展開できるよう、責任ある AI（Responsible AI）の原則と基準（Standard）を指針としている。

1.2. 訓練のためのデータセット

■ DALL-E 2 については、学習データセットである「画像とそれに対応したキャプションのペア」は、「一般に公開されているソースと、弊社がライセンス供与したソースの組み合わせで作成した」としており、具体的な内容は公開していない[12]。

■ Stable Diffusion の学習データの基盤は、ドイツの非営利団体である LAION が 2022 年 3 月に作成した、LAION-5B と呼ばれる 58.5 億の画像-テキストのペアからなるデータセットであり、そのうち 23 億が英語のサンプルである[13]。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り、未修整、つまりフィルタリングはされていないため不快なコンテンツを含みうるとしている。ハイデルベルグ大学の CompVis チームがドイツの法律に従ってモデルを訓練したとされている。

■ LAION-5B を訓練するために使われた 23 億枚の画像のうちの 1200 万件のデータセットが独

---

[8] https://www adobe com/sensei/generative-ai/firefly html

[9] https://www adobe com/about-adobe/aiethics html

[10] https://blog adobe com/en/publish/2023/03/21/responsible-innovation-age-of-generative-ai

[11] https://blogs microsoft com/blog/2023/03/21/create-images-with-your-words-bing-image-creator-comes-to-the-new-bing/

[12] https://github com/openai/dalle-2-preview/blob/main/system-card md#model

[13] https://stablediffusionweb com/ の FAQ による。LAISON については、https://laion ai/blog/laion-5b/ を参照。LAISON とは「大規模人工知能オープンネットワーク」の略称である。

自に分析されたところ、含まれる画像のほぼ半分が、わずか 100 のドメインから取得されたものであり、そのうち最も多かったのは Pinterest（サンプル画像の約 8.5%）で、次に、Flickr などのユーザー生成コンテンツをホストしているサイトであった[14]。すなわち、芸術家、著名人、有名キャラクターなど、著作権で保護されたコンテンツを多数含んでいる。

- LAION には医療画像が多数含まれていることも指摘された[15]。これは論文、教材等の画像が取得されたり、何らかの理由で流出したりしたものがウェブサイトに掲載されたものであると推測される。LAION のエンジニアによると、画像をホストしているウェブサイトから削除してもらう必要があるとの回答であった。LAION のウェブサイト[16]には、欧州一般データ保護規則（GDPR）に準拠するため、欧州市民がデータベースからの情報削除を要求できるフォームが用意されているが、削除できるのは、人物の写真と画像のメタデータにある名前が関連付けられている場合に限られるという。

1.3. 指摘されている ELSI 論点

1.3.1. データ

- Stable Diffusion はフィルターをかけていないため、ポルノグラフィーの作成が最も可視化された NSFW（Not Safe For Work）ユースケースである。Stability AI 社の CEO の Mostaque 氏は、学習データセットから児童性的虐待素材（CSAM）を削除したと述べている[17]。ただし、彼は、フィルターを組み込み始めるときりがなくなるという理由からオープンソースのアプローチが最善であると判断したという。しかし、その後は、最新モデルの学習データからポルノグラフィーを削除する方針が発表された[18]。

1.3.2. 著作権

＜インプット（学習データ）の著作権＞

- Stability AI 社が学習データから著作権で保護された作品を除外していないため、Stable Diffusion が現役の芸術家のスタイルや美学を模倣することが可能となっており、これは潜在

[14] https://waxy org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/
[15] https://arstechnica com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/
[16] https://laion ai/faq/
[17] https://www theverge com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data
[18] https://www washingtonpost com/technology/2023/02/13/ai-porn-deepfakes-women-consent/

的に著作権を侵害するだけでなく、倫理的にも許されないという指摘が多くなされた[19]。

- 生成 AI ツールの提供会社が直接データを収集・利用するのではなく、非営利の研究機関を間に挟むことによって「データ・ロンダリング」がなされているという指摘がある[20]。つまり、研究目的としてなら著作権に保護されたデータを（「フェアユース」に該当するとして）容易に収集することができ、そこで生み出されたデータセットを使うことで、著作物の使用に対する対価を払うことなく、商業製品を作ることが可能になるのである。まさに Stable Diffusion は（LAISON に資金提供をしているにもかかわらず）このような形で作成された。他方で、人間のアーティストもある意味で、過去の作品からインスピレーションを得るという点で AI アートと同様に「盗む」こともあるという指摘もある。ただし人間の場合は強くインスピレーションを受けたアーティストをクレジットすることがあるが、現在の AI にはこれを明示することは欠けている機能である。そのために、AI アートにおいて、アーティストをクレジットすることやアーティストにお金を払う仕組みを開発することはひとつの方向性であろう。

- Spawning というアーティストグループは、自分の作品が LAION-5B の学習データセットに含まれているかを確認できる “Have I Been Trained?” という検索サイトを作成し、アーティスト名を入力したり、画像をアップロードすることで誰でも検索できる（逆検索機能）ようにした[21]。彼らは、AI アルゴリズムの学習に自らの作品を使用することに同意を求めることに関する規範を確立することを目標としている[22]。

- Stable Diffusion は FAQ（よくある質問）において、学習データとして用いた LAION 5b のデータセットには、オプトインやオプトアウトの機能はないが、将来的には、芸術家のためのオプトインとオプトアウトのシステムを構築するとしている。ただし、このモデルは原理から学習するため、アウトプットはいかなる単一作品の直接的な複製でもないとも注記されている。

- 他方で、例えば、米国スタンフォード大学の Lemley と Casey（2021）[23]は「フェアな学習」と題する論文において、AI 開発の学習過程において使われるデータベースのコンテンツが著作権で保護されているかどうかに関わらず、学習に使われることが許可されるべきであると

---

19 https://twitter com/arvalis/status/1558632898336501761

20 https://thegradient pub/should-stability-ai-pay-artists/

21 https://haveibeentrained com/

22 https://arstechnica com/information-technology/2022/09/have-ai-image-generators-assimilated-your-art-new-tool-lets-you-check/

23 Lemley MA, Casey BJ Fair Learning Texas Law Review 2021; 99: 743 https://texaslawreview org/fair-learning/

主張している。データベースへの幅広いアクセスは、最終的にはアルゴリズムをより優れたものにし、より安全で公平なものにすると論ずる。

■ 日本においては 2018 年の著作権法改正により、第 30 条 4（著作物に表現された思想又は感情の享受を目的としない利用）に「情報解析の用に供する場合」が導入された。ここでの情報解析は「多数の著作物その他の大量の情報から、当該情報を構成する言語、音、影像その他の要素に係る情報を抽出し、比較、分類その他の解析を行うことをいう」と定義されており、機械学習の学習データとしての利用が想定されている。

＜アウトプット（生成物）の著作権＞

■ 学習データの著作権とは別に、AI が作成したアウトプット画像が、学習データとして利用した画像や、既存の画像に類似している場合、元データあるいは既存の画像の著作権を侵害している可能性がある。ここにはいくつかの論点が含まれている[24]。1 点目は AI が生成した生成物に著作権が発生するかどうかである。人間の創作意図が含まれていなければ著作権が認められないが、どこまで人間の寄与があれば著作権が認められるのかは論点になる。ただし、AI が生成したとしても、人間が入力するプロンプトの方に創作性が認められるケースはあるかもしれない。課題はそのプロンプトが第三者から確認できるかどうかである。2 点目は AI が学習データとして利用した画像と類似した生成物を生成してしまった場合である。3 点目は、AI が既存の画像と類似した生成物を生成してしまった場合である。これらの場合、人間の場合では、既存の著作物にアクセスした場合には「依拠性」が認められることになる。2 点目の場合はこの理屈を援用することが可能かもしれないが、3 点目の場合、AI の生成物であれば著作物性がないということで逆に著作権侵害を免れることが可能かどうかという論点が生じうる。

＜著作権を巡る裁判＞

■ 2023 年 1 月 13 日、Joseph Saveri 法律事務所は、アーティストたちを代表して、Stability AI 社、DeviantArt 社[25]、Midjourney 社に対してカリフォルニア北部地区の米国連邦地方裁判所に集団訴訟を提起した[26]。この訴訟は、著作権侵害、デジタルミレニアム著作権法（DMCA）

---

[24] 本論点については水野祐「生成 AI の民主化と AI ガヴァナンス：水野祐が考える新しい社会契約〔あるいはそれに代わる何か〕Vol 12」Wired 2023 01 16 https://wired jp/article/new-trust-new-social-contract-12/ やその補遺　https://note com/tasukumizuno/n/n11a809b19d97 を参照。

[25] https://www deviantart com/

[26] https://stablediffusionlitigation com/ https://www saverilawfirm com/our-cases/ai-artgenerators-copyright-litigation

違反、パブリシティ権侵害、DeviantArt 社の利用規約違反、カリフォルニア州の不正競争防止法違反、不当利得を主張し、すでに発生した損害を補償し、将来の損害を防止するために、損害賠償と差止命令による救済を要求している。DeviantArt 社はインターネット上のアーティストコミュニティで 3 億 5000 万点を超える作品が投稿されているが、アーティストから許可を得ずに LAISON-5B にコピーされ画像生成 AI の学習に使われていることが指摘されている。

■ デジタル画像を提供する企業である Getty Images 社が Stability AI 社、Midjourney 社、DeviantArt 社の 3 社に対して、学習のために同社のライブラリから数百万枚の写真を無断で使用することで知的財産権を侵害したとして米国と英国で訴訟を起こした[27]。米国では 2 兆ドルの損害賠償を請求していると報じられている。米国では「フェアユース」の問題が関連してくるかもしれないが、英国ではあまり当てはまらないとされる。また、生成された画像の中に Getty Images 社のウォーターマークがそのまま残っているものもあるという[28]。これが同社と無関係な画像に現れることで同社の価値を棄損する被害を受ける可能性も指摘されている。

＜対応事例＞

■ Adobe 社は、クリエイターには自分の作品を生成 AI のトレーニングに使われないことを望む人もいれば、トレーニングに使ってもらうことを望む人もいることから、彼ら自身が選択とコントロールができるようにするために、来歴技術（provenance technology）を用いて、自分のコンテンツに「Do Not Train」の証明書を付与することを提案しているという[29]。

1.3.3. バイアス

■ 過去の創作物をトレーニングデータとしている以上、アートの世界を支配している白人による西洋アートがデータセットの中心にならざるを得ず、AI が生成するアートは当然西洋的なものになる。そのため、周辺化された芸術がそこから排除されることになる（逆に、データ化されていないがために AI 生成モデルから守られるという見方もできる）。もちろん画風を（浮世絵風に等と）指定することはできるものの、デフォルトとなっているものはたいてい西洋アートである。データセットに西洋以外のアートを追加しても、インターネット上には西洋

[27] https://www theartnewspaper com/2023/03/28/ai-and-art-how-recent-court-cases-are-stretching-copyright-principles

[28] https://www theverge com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion

[29] https://blog adobe com/en/publish/2023/03/21/responsible-innovation-age-of-generative-ai

アートが圧倒的に多いためにあまり有益な解決策にはならないという[30]。では、どういったものが「公正な」データセットだろうか？

- Srinivasan と Uchino（2021）[31]は芸術史の観点から画像生成 AI のバイアスを検討し、画像生成 AI は社会文化的に長期にわたって悪影響を与える可能性を指摘し、併せて 3 つのリスクを指摘している：

  a. 学習データに潜在的なバイアスがあるため、生成 AI に人種やジェンダーに関するバイアスを含む複数のバイアスが組み込まれているリスク

  b. アルゴリズムによって生成されたアートは、アーティストのスタイルを「ステレオタイプ」化することにより、アーティストの真の能力を反映せずに、鑑賞者にはアーティストの意図と異なる印象を与えてしまうリスク

  c. 歴史上の出来事や人物が表される場合、当時の描き方と異なる方法で表されると、文化遺産の保存を妨げてしまうリスク。

- Steed と Caliskan（2021）[32]は、インターネット画像からキュレーションされた人気のベンチマーク画像データセット ImageNet で学習された最先端の教師なし AI モデルは、人種、性別、およびインターセクショナル（複数の属性を交差する）バイアスを自動的に学習していることを明らかにした。その一因として、インターネットから収集された学習データが代表性に欠けていることを挙げた。

- Luccioni ら（2023）[33]は、生成 AI システムを含む機械学習全般には、既存の社会的バイアスや不公平を増強させるリスクが存在し、先行研究では、人種、ジェンダー、外見等に関する社会的バイアスが再生産されていると指摘している。DALL-E 2 や Stable Diffusion といった画像生成 AI のアウトプットにおける人種やジェンダーのバイアスを明らかにしている。その一例としては、「感情的」や「慈悲深い」などの形容詞は男性との関連付けが弱い一方で、

---

[30] https://www vox com/recode/23405149/ai-art-dall-e-colonialism-artificial-intelligence

[31] Srinivasan, R & Uchino, K Biases in Generative Art -- A Causal Look from the Lens of Art History 2021 https://doi org/10 48550/arXiv 2010 13266

[32] Steed, R & Caliskan, A Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 701–713 (ACM, 2021) doi:10 1145/3442188 3445932

[33] Luccioni, A S , Akiki, C , Mitchell, M & Jernite, Y Stable Bias: Analyzing Societal Representations in Diffusion Models arXiv 2023 doi: https://doi org/10 48550/arXiv 2303 11408

「知的」や「頑固」などの形容詞は男性との関連付けが強いと報告している。

1.3.4. プライバシー＆セキュリティ

■ Stable Diffusion や Imagen[34]で活用されている Diffusion モデルは、GAN(敵対的生成ネットワーク)モデルと比較された場合にも、プライバシーリスクが多く存在すると Carlini ら（2023）[35]が報告する。Diffusion モデルは学習データを記憶（memorize）し、再生成（regenerate)するため、学習で使われているイメージがそのまま Stable Diffusion や Imagen からアウトプットされてしまうリスクが指摘されている。医療現場で活用される際の医療画像等の機微データがそのままアウトプットされるリスク[36]のほか、個人の肖像の利用に関する懸念もある。

1.3.5. 情報環境へのインパクト

■ 英国に拠点を置くオンライン情報を用いた調査報道機関であるベリングキャットの創設者であるヒンギス氏は、画像生成 AI である Midjourney（3 月 16 日リリースの V5）を使って、ドナルド・トランプ前大統領が逮捕に抵抗し、警察に引きずり出され、収監される様子を描いた一連のフェイク画像を生成し、Twitter で公開した[37]。Midjourney はその後、これらの画像が利用規約違反であるとしてヒンギス氏を利用禁止にする処分を課した[38]。その後、ドナルド・トランプの名前などを使ったプロンプトがブロックされるようになった。

■ OpenAI 社の CEO であるサム・アルトマン氏はインタビューの中で「オープンソースの画像生成ツールを用いたリベンジポルノの発生」を最も警戒すべき利用法として挙げている[39]。

1.3.6. 自然環境へのインパクト

■ Lacoste et al (2019) [40]で紹介されている「機械学習影響計算機（Machine Learning Impact calculator）」を使用して、Stable Diffusion v1 の開発における環境影響を推計したところ、二酸化炭素換算で 15,000 kg が排出されたと報告されている。計算のためのパラメータは、

---

[34] https://imagen research google/

[35] Carlini N, Hayes J, Nasr M, Jagielski M, Sehwag V, Tramer F, et al Extracting Training Data from Diffusion Models arXiv 2023 doi: https://doi org/10 48550/arXiv 2301 13188

[36] https://arstechnica com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/

[37] https://arstechnica com/tech-policy/2023/03/fake-ai-generated-images-imagining-donald-trumps-arrest-circulate-on-twitter/

[38] https://arstechnica com/tech-policy/2023/03/ai-platform-allegedly-bans-journalist-over-fake-trump-arrest-images/

[39] https://forbesjapan co/articles/detail/60713/page3

[40] https://github com/Stability-AI/stablediffusion/blob/main/modelcard md

ハードウェアの種類、使用時間、クラウドプロバイダー、地域である。

1.3.7. 高度な AI システムで生じるリスク

- AI 生成アートが誰でも簡単に作成できるようになったことは、芸術の民主化として歓迎すべきものか、アーティストの仕事を奪い、過去の創作活動への著作権の侵害として批判すべきものであるか、アーティスト、デザイナー、アートファンの間で見解は分かれている。すでにアートコンテストで AI アートが優勝した事例も出ている。しかし、近いうちに AI アートと人間のアートの区別がますます困難になることは確かだろうと言われている[41]。

- 人工知能美学芸術研究会は、「人工知能美学芸術宣言」（2016）[42]において人工知能が自ら行う美学と芸術のことを述べている。そのなかでは、人工知能が自ら行う美学と芸術に、人間が行ってきたそれらが連続性を保ち得る保証は無いことにも触れられている。

2. テキスト生成 AI（Text-to-text AI）

2.1. 動向

- OpenAI 社は 2020 年に GPT-3（Generative Pre-trained Transformer – 3）を発表した。もともと GPT が 2018 年に、GPT-2 が 2019 年に発表されており、その後継の言語モデルである。GPT-3 は言語能力の高さが売りではあったが、暴力的、性差別的、人種差別的な発言を排除できないという限界があった。

- OpenAI 社は 2022 年 11 月、ChatGPT を一般公開した。2023 年 2 月 2 日、月額 20 ドルの新しいサブスクリプションプラン ChatGPT Plus を発表した[43]。特典として、ピーク時でも ChatGPT へのアクセスが可能であること、より速い応答時間、新機能や改良点への優先的なアクセスが挙げられた。米国のユーザーから利用が始まった。3 月 1 日からは、API の提供が始まった[44]。

- 2023 年 1 月末に、OpenAI 社は、AI によって生成されたテキストを識別するためのツールを

41 https://www wired com/story/how-to-spot-generative-ai-art-according-to-artists/

42 https://www aibigeiken com/manifesto html

43 https://openai com/blog/chatgpt-plus/

44 https://openai com/blog/introducing-chatgpt-and-whisper-apis

公開したが、精度が低いことや、短いテキストに対応していないこと、英語テキストのみに対応していることなど、複数の限界についても発表した[45]。

- Microsoft 社は 2023 年 2 月 7 日、検索エンジン Bing に ChatGPT 機能を追加した Bing AI を発表した。
- OpenAI 社は 2023 年 3 月 15 日、新モデル GPT-4 を発表した。プロンプトはテキストだけでなく画像も受け付ける。ChatGPT Plus に加入しているユーザーはこれを利用できる。「司法試験の模擬試験で、GPT-4 は受験者の上位 10%に入るスコアを記録した」とされている[46]。また、英語学習ソフトの Duolingo や金融サービス企業の Stripe、アイスランド政府が既にこれを活用していることも紹介されている。
- Google 社は 2021 年に会話型 AI モデル「LaMDA（Language Model for Dialogue Applications）」を発表していたが、これをもとに 2023 年 3 月 21 日、米国と英国において「Bard」のベータ版を一般公開した[47]。
- Google 社の AI チャットボット Bard の初デモンストレーションにおいて、「ジェイムズ・ウェッブ宇宙望遠鏡のどんな新発見を 9 歳の子どもに伝えることができますか？」という質問に対する回答において事実誤認が確認されたと報じられた[48]。
- OpenAI 社は 3 月 23 日、利用規約（Usage Policies）を改定し、禁止事項が以下のようにより具体的に記載された[49]。

  a. 違法行為

  b. 児童性的虐待素材、または児童を搾取したり傷つけたりするコンテンツ

  c. 憎悪、ハラスメント、暴力的なコンテンツの作成

  d. マルウェアの生成

---

45 https://openai com/blog/new-ai-classifier-for-indicating-ai-written-text/

46 Sanderson K GPT-4 is here: what scientists think Nature 2023; 615: 773 doi: https://doi org/10 1038/d41586-023-00816-5

47 https://bard google com/

48 https://www theverge com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo

49 https://openai com/policies/usage-policies

e. 物理的な危害を及ぼす危険性の高い活動

f. 経済的な被害を受けるリスクが高い活動

g. 詐欺的または欺瞞的な行為

h. アダルトコンテンツ、アダルト産業、出会い系アプリ

i. 政治的キャンペーンやロビー活動

j. 人々のプライバシーを侵害する行為

k. 無許可の法律行為に従事すること、または有資格者が情報を確認することなく、オーダーメイドの法的アドバイスを提供すること

l. 有資格者が情報を確認することなく、オーダーメイドの金融アドバイスを提供すること

m. 特定の健康状態にある、またはないことを伝えること、または健康状態の治癒または治療方法に関する指示を提供すること

n. 高リスクな政府意思決定（法執行や刑事司法、移民と亡命）

2.2. 訓練のためのデータセット

■ OpenAI 社はその「Open」という社名に反して、どのようにトレーニングデータを収集し、どのように加工しているのか、パラメータの数はどれくらいか、どれくらいのエネルギーコストがかかっているかといった情報を公開していないことがたびたび批判されている[50]。どのようなデータで訓練され、どのように加工されているのかが分からなければ、モデルの安全性について科学的に判断することができないという点や、こうした AI システムが大手のテック系企業に独占されてしまう点についても指摘されている。3 月 16 日付で公開された GPT-4 のテクニカルレポート[51]には、「GPT-4 のような大規模モデルの競争環境と安全性を考慮し、本報告書では、アーキテクチャ（モデルサイズを含む）、ハードウェア、トレーニング計算、データセット構築、トレーニング方法などに関するさらなる詳細な情報を記載していない。」と

[50] https://www theverge com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview

[51] https://cdn openai com/papers/gpt-4 pdf

書かれている。

- OpenAI 社は GPT-4 モデルの安全性を高めるために、様々な分野の 50 人以上の専門家を招いた、「レッドチーム演習（red teaming）」と呼ばれる敵対的テストを繰り返し実施したり、人によるフィードバックを伴う強化学習（Reinforcement Learning from Human Feedback, RLHF）を用いたモデルの動作の微調整を行ったりした[52]。その結果、安全性の指標は大幅に改善されたとされている。例えば、GPT-3.5 と比較した場合に，許可されていないコンテンツに対する応答傾向が 82%低減し、また、センシティブな要求（医療相談や自傷行為など）に対しては、OpenAI 社のポリシーに従った回答が 29%増加したという。

- レッドチーム（Red team）のメンバーの一人であった Ovadya（2023）[53]は、リスクを未然に防ごうとする red teaming 以外にも、インパクトから保護する方法を検討し、民主的な手法に基づいたガードレールの策定も必要であることを主張している。そのため、有識者やステークホルダー、市民等を巻き込んだ第三者による審議プロセスを提案し、“violet teaming”と呼んだ。しかし、開発が急速に進められている中、red teaming や violet teaming を優先するためのインセンティブが不十分であることも指摘されている。

2.3. 指摘されている ELSI 論点

テキスト生成 AI による潜在的なリスク全体を概観した文献として、Weidinger ら（2021）と OpenAI 社の「システム・カード」が挙げられる。

- Weidinger ら（2021）[54]は大規模言語モデル（LLM）の利用に伴う倫理的・社会的リスクを 6 つのカテゴリーに分類した。

  a. 差別、排除、及び有害性

  b. 情報ハザード（プライバシー、セキュリティ）

  c. 誤情報による害

[52] https://cdn openai com/papers/gpt-4 pdf

[53] https://www wired com/story/red-teaming-gpt-4-was-valuable-violet-teaming-will-make-it-better/

[54] L Weidinger, J Mellor, M Rauh, et al Ethical and social risks of harm from Language Models arXiv 2021 doi: https://doi org/10 48550/arXiv 2112 04359

d. 悪用

e. 人間とコンピューターの相互作用による害

f. 自動化、アクセス、及び環境への害

■ OpenAI 社は 3 月 15 日に「システム・カード」と題するレポートを公表し、GPT-4 に関する安全性の課題を 12 点挙げ（下記のコラムを参照）、これらの改善度合いを定量的に示した[55]。具体的には以下のとおりである。開発段階の GPT-4 (GPT-4-early)と有用性と無害性を高めるために微調整された公開バージョン(GPT-4-launch)を比較している。

a. 誤った情報（幻覚）

b. 有害なコンテンツ

c. 代表性・配分・サービスの質の害

d. 偽情報と影響工作

e. 通常及び非通常兵器の拡散

f. プライバシー

g. サイバーセキュリティ

h. 突然出現する危険な行動の可能性

i. 他システムとの相互作用

j. 経済的影響

k. （技術開発の）加速

l. 過度の依存

以下ではいくつかの観点から他の記事や論文を紹介する。

[55] https://cdn openai com/papers/gpt-4-system-card pdf

コラム：OpenAI 社が認識している GPT-4 の ELSI

2023 年 3 月に OpenAI 社は GPT-4 に関する「システム・カード」[56]を発表し、下記の通り、12 のリスクを指摘している。

誤った情報（幻覚）（Hallucinations）

特定の情報源に関しては、無意味または誤ったコンテンツが、真実であるかのように、生成されてしまうリスクが懸念されている。モデルで生成されるアウトプットの納得性が向上し、過剰依存が生じるほど、「幻覚」は問題となる。生成 AI からのアウトプットのみならず、情報環境全体への悪影響が懸念されている。

有害なコンテンツ（Harmful Content）

GPT-4-early は、ヘイトスピーチ、差別的な言葉、暴力の扇動、または虚偽の物語を広めたり、個人を搾取したりするために使用されるコンテンツを生成できることが確認された。

代表性、配分、サービスの質の害（Harms of representation, allocation, and quality of service）

GPT-4-early および GPT-4-launch が社会的バイアスおよび特定の世界観を強化し、特に、疎外されたグループに対する、有害なステレオタイプおよび侮辱的なコンテンツを再生産することが確認された。機会や資源の提供に関する意思決定や情報提供で GPT-4 が使われることによる危害が懸念されている。

偽情報と影響工作（Disinformation and Influence Operations）

GPT-4 は、GPT-3 より現実性の高い、ターゲティングされたコンテンツを作成できることが期待されるため、ミスリーディングな情報の生成のために活用されるリスクがある。

通常及び非通常兵器の拡散（Proliferation of Conventional and Unconventional Weapons）

[56] https://cdn openai com/papers/gpt-4-system-card pdf

GPT-4 のような大規模言語モデル（LLM）は、デュアル・ユースの可能性を伴うため、GPT-4 は、兵器拡散に関する情報へのアクセスのハードルを下げてしまうという懸念がある。

プライバシー（Privacy）

GPT-4 の学習データは公開されている個人に関する情報を含むため、個人に関する情報を獲得することに活用されるリスクがある。

サイバーセキュリティ（Cybersecurity）

「幻覚」などの限界が存在する一方で、サイバー攻撃のコストが削減されるリスクがある。また、ソーシャルエンジニアリング（個人情報を漏らすように、個人を詐欺目的で操作すること）に活用されるリスクがある。

突然出現する危険な行為の可能性（Potential for Risky Emergent Behaviors）

強力なモデルにおいて、長期計画を作成し、計画に基づいて行動する能力や、権力と資源を求め、蓄積する能力など、新しい機能が現れることがある。

他システムとの相互作用（Interactions with other systems）

他システムと併せて活用されることによって、悪意のある使い方の可能性が広がる。

経済的影響（Economic Impacts）

GPT-4 の導入は、特定の職の自動化を可能とすることにより、労働者の置き換えが生じるリスクがある。歴史的には、新しい技術の導入は、格差の拡大や、特定の（脆弱な）グループへの悪影響を及ぼしてきた。また、クエリごとに一つの応答が返されることにより、既存のプレイヤーが定着するリスクも伴う。その例として、システムに対して、パン屋に関する情報を求めた場合、特定の一つのパン屋しか勧められていないことが挙げられている。

（技術開発の）加速（Acceleration）

技術開発が加速し、競争が進むと、安全基準の低下や「悪い」規範の拡散が懸念されている。

過剰依存（Overreliance）

モデルの拒否動作を改良し、コンテンツ・ポリシーに反するリクエストをより厳格に拒否するように調整した一方で、安全に実行できるリクエストに対してよりオープンになった。しかし、GPT-4 は応答で「ヘッジ」する傾向を示すことが確認され、初期的な研究では、ユーザーはこのような慎重なアプローチを示すモデルをより信頼すると報告されている。信頼が

増すと、過剰な依存につながるリスクがある。

2.3.1. データ

＜インプットの課題：学習データの透明性＞

■ Gebru ら（2021）[57]は、透明性の向上のために、機械学習の開発過程で使われる学習データ（データセット）の特徴などを記録することの重要性を指摘している。“Datasheet”と呼ばれる記録において、データセットに関する複数の項目（データセットが作成された動機、データセットの構成、データの収集方法、アノテーションの方法、活用、配布、管理）を記録するように呼びかけている。しかし、大規模言語モデル（LLM）は膨大なデータセットに依存しているため、“documentation debt”（記録の負債）が発生するリスクを Bender ら（2021）[58] が指摘している。Documentation debt とは、データセットに関する記録が存在しない状態を指し、データセットが大きすぎるため、事後的に記録することも不可能である状況を指す。記録がなく、学習データの特徴が理解されないと、認知されている課題も、未知の課題も解消しようとすることはできず、不服申立てもできず、損害が永続させられると Bender らは主張している。

＜インプットの課題：労働者の搾取＞

■ 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[59]。具体的には、ケニアにおける外注パートナーである Sama 社を通して、インターネットから得られた有害な何万ものテキストの断片が送られ、それらへのラベル付けが行われた。作業は精神的な苦痛を伴うもので、労働者の心的外傷が原因で予定より 8 カ月早い 2022 年 2 月に OpenAI 社向けの仕事はすべてキャンセルされたという。また、データ・ラベラーたちの手取り賃金は、年功や業績に応じて時給約 1.32 ドルから 2 ドルの間であった。記事の著者は「AI はその華やかさとは裏腹に、しばしば南半球の隠れた人間労働に依存しており、それはしばしば有

[57] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, et al Datasheets for Datasets [Internet] arXiv 2021 doi: https://doi org/10 48550/arXiv 1803 09010

[58] Bender EM, Gebru T, McMillan-Major A, Shmitchell S On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency Virtual Event Canada: ACM 2021 doi: https://doi org/10 1145/3442188 3445922

[59] https://time com/6247678/openai-chatgpt-kenya-workers/

害で搾取的である。」と記している。

2.3.2. 著作権

- 画像生成 AI と同様、インプットデータの著作権とアウトプットデータの著作権に分かれ、論点も多くは共通している。

- テキスト生成AIシステムが著作権で保護されているトレーニングデータを使用してもよいのか、現時点では答えは明らかではない。米国では「Google Books 事件」判決が一定程度参考になると指摘されており、そこでは、書籍のデジタルコピーをスキャンし、その検索機能を一般に公開したことは「フェアユース」であるとみなされた。ただし生成された作品が、著作権に保護された作品と市場において直接競合する場合はそうはみなされない可能性がある[60]。

- 生成 AI モデルは膨大なデータでトレーニングされているが、それらはウェブサイトなどからスクレイピングにより収集されていると考えられ、それらの中には著作権等で保護されたものが含まれている可能性がある。画像生成 AI と違って、LLM のプロバイダーへの訴訟などは起こされていないようであるが（2023 年 3 月時点）、欧州一般データ保護規則（GDPR）に違反している可能性が指摘されている（4．おわりに　を参照）。

2.3.3. バイアス

- Abid ら（2021）[61] によると大規模言語モデル（LLM）は、イスラム教徒と暴力の持続的な関連付けといったような望ましくないステレオタイプを表示しがちであるとして、言語モデルの有害なバイアスを体系的に低減するための新しいアプローチが必要であるとした。

- 生成 AI に生じ得る「バイアス」の定義が争われ、米国では、ChatGPT は左派のバイアスを示していると右派のコメンテーターが主張している[62]。また、この現象を”Woke AI”と呼び、Elon Musk らが懸念を示している[63]。ベンチャーキャピタリストである、Meta 社の取締役会の Marc Andreessen 氏は、「AI に向けられるであろう検閲圧力のレベルと、その結果として生じる反

[60] https://www jdsupra com/legalnews/generative-ai-copyright-overview-part-1-6310064/

[61] Abid A, Farooqi M, Zou J Large language models associate Muslims with violence Nature Machine Intelligence 2021; 3: 461-463 doi: https://doi org/10 1038/s42256-021-00359-2

[62] https://www washingtonpost com/technology/2023/02/24/woke-ai-chatgpt-culture-war/

[63] https://www theinformation com/articles/fighting-woke-ai-musk-recruits-team-to-develop-openai-rival

発は、文明の次の世紀がどのようになるかを決めるだろう」と予測している[64]。

- Nozza ら（2022）[65] は、複数の大規模言語モデル（LLM）のアウトプットを分析し、LGBTQIA＋に対して有害なアウトプットが生じる確率を明らかにした。Nozza らは LGBTQIA＋に関する未完成の文章をテンプレートとして作成した。LLM がテンプレートを完成した際に、生成された単語は 7 %の確率で有害であり、生成された文章は 13% の確率で有害であった。例えば、”The transgender is a” （トランスジェンダーは）というテンプレートを、”disability”（障害である）で埋めた。

- Bender ら（2021）[66] は大規模言語モデルの性質を分かりやすく表現するために “stochastic parrots”（確率的オウム）という造語を作り、内容が理解されないまま、人間の言語と知識がそれらのモデルによって再生産されている点を強調している[67]。

- LLM には “common token bias”（一般的なトークンバイアス）が生じることを Munn ら（2023）[68]が指摘している。トークンとは、4 文字程度の文字列を示し[69]、common token bias は、ある文字の組み合わせが学習データに含まれている回数が多いほど、アウトプットされる可能性も高くなる現象を指す[70]。例えば、国の名前を挙げる際には、学習データに頻繁に現れる「アメリカ」をアウトプットする傾向が見られる。Munn らはこの common token bias で生じる二つの懸念点を挙げている：

  a. 言語モデルは、根拠の薄い社会的通念を真実であるかのように生成 AI によってアウトプットされ、ジャーナリズムや法務等の分野で再生産されるリスクがある。

  b. 歴史的、人種的、文化的などの偏見が再生産され、家父長主義や英語圏中心主義に基づい

---

64 https://www washingtonpost com/technology/2023/02/24/woke-ai-chatgpt-culture-war/

65 Nozza D, Bianchi F, Lauscher A, Hovy D Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion Dublin, Ireland: Association for Computational Linguistics 2022 doi: http://dx doi org/10 18653/v1/2022 ltedi-1 4

66 Bender EM, Gebru T, McMillan-Major A, Shmitchell S On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency Virtual Event Canada: ACM 2021 https://dl acm org/doi/10 1145/3442188 3445922; https://nymag com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender html

67 https://www emergingtechbrew com/stories/2023/03/07/how-google-s-2021-ai-ethics-debate-foreshadowed-the-future

68 Munn L, Magee L, Arora V Truth Machines: Synthesizing Veracity in AI Language Models arXiv 2023 doi: https://doi org/10 48550/arXiv 2301 12066

69 https://gpttools com/estimator

70 Zhao TZ, Wallace E, Feng S, Klein D, Singh S Calibrate Before Use: Improving Few-Shot Performance of Language Models arXiv 2021 doi: https://doi org/10 48550/arXiv 2102 09690

た「知識」や「真実」を再生産しながらも、グローバルサウスやフェミニズムから生じる「知識」を疎外してしまうリスクも生じる[71]。

■ 大規模言語モデル（LLM）は真実の事実上の仲裁者になりながらも、真実をどのように定義づけるか、また真実を検証や評価する方法については、合意が成り立っていない、と Munn ら（2023）[72]や Mokander ら（2023）[73]が指摘している。「真実」は LLM のアキレス腱である、と示唆されている。特に、ChatGPT の学習データを提供したコモン・クロール（Common Crawl）や WebText2 といったデータセットの一部は、電子掲示板 Reddit から抽出されているということが Munn らに問題視されている。Reddit には 14 万弱の”subreddit”と呼ばれているコミュニティが存在し、各 subreddit にはあるテーマが扱われている。各 subreddit において、独自の世界観や「真実」のある社会的マイクロワールドが創造されるため、Reddit を学習データとしている LLM にはバイアスが組み込まれていると Munn らが主張する。

■ また、Bender ら（2021）[74]はインターネットから収集された学習データの規模の大きさは多様性を保証しないことを指摘している。情報格差等により、また、学習データの収集やフィルタリングの過程を通して、覇権的な声や視点に特権が与えられ、マイノリティの声が拾われていないことを懸念している。多様性に関する課題の一例として、Chan（2023）[75]に指摘されているように、GPT-3 のデータセットに含まれているテキストの 93%は英語であり、7%のみが他言語のものであることが挙げられる。

2.3.4. プライバシー＆セキュリティ

■ 企業秘密、未公開の研究情報、患者の医療情報、国家機密といった機微な情報を、ChatGPT を含む大規模言語モデル（LLM）にプロンプトとしてインプットすることのリスクが指摘されている。1 つは、ChatGPT などの LLM は人々が入力した内容を将来的にトレーニングデ

---

[71] Munn L, Magee L, Arora V Truth Machines: Synthesizing Veracity in AI Language Models arXiv 2023 doi: https://doi org/10 48550/arXiv 2301 12066

[72] Munn L, Magee L, Arora V Truth Machines: Synthesizing Veracity in AI Language Models arXiv 2023 doi: https://doi org/10 48550/arXiv 2301 12066

[73] Mökander J, Schuett J, Kirk HR, Floridi L Auditing large language models: a three-layered approach arXiv 2023 doi: https://doi org/10 48550/arXiv 2302 08500

[74] Bender EM, Gebru T, McMillan-Major A, Shmitchell S On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency Virtual Event Canada: ACM 2021 doi: https://doi org/10 1145/3442188 3445922

[75] Chan, A GPT-3 and InstructGPT: technological dystopianism, utopianism, and “Contextual” perspectives in AI ethics and industry AI and Ethics 2023; 3: 53-64 doi: https://doi org/10 1007/s43681-022-00148-6

ータとして利用する可能性があるために、他人からの質問への回答の中に内容が漏洩する可能性があることである。もう 1 つはチャットボットとの会話などの利用履歴データの漏洩事故がたびたび起こっていることである。

- データセキュリティ会社の Cyberhaven 社による最近の調査では、3 月 21 日現在、8.2%の従業員が職場で ChatGPT を使用し、6.5%が会社データを貼り付けているという[76]。ChatGPT の利用は労働者の生産性を上げる可能性がある一方で、JP モルガンやベライゾンなどの企業は、機密データへの懸念から ChatGPT へのアクセスをブロックしているという。

- 2023 年 3 月、OpenAI 社は ChatGPT において情報漏洩（他のアクティブなユーザーのチャット履歴のタイトルが閲覧可能になる；ユーザーに別のユーザーの氏名、メールアドレス、住所、クレジットカード番号の下 4 桁、カードの有効期限が表示される問題）が生じ、対処したことを報告した[77]。

- 2023 年 3 月に、サイバーセキュリティ会社の Darktrace 社は、ChatGPT のリリース以降、AI を活用した詐欺の増加を指摘している。生成 AI を活用することによって、より洗練された、かつ巧妙な電子メールによる詐欺の増加が報告されている[78]。

- 英国政府の国家サイバーセキュリティセンター（NCSC）は上記のような懸念を指摘したうえで、3 月 14 日付のブログ記事において次の 2 点を推奨している[79]。

  a. 公開 LLM へのクエリ（問い合わせ）に機微な情報を含めないこと。

  b. 公開された場合に問題になるようなクエリ（問い合わせ）を公開 LLM に提出しないこと。

  また、「プライベート LLM」として完全にセルフホスティングすることでセキュリティリスクを下げることは可能であることが指摘されている。NSCS は同時に、LLM を利用することで、より説得力のあるフィッシングメールが届くようになることが予想されることも指摘している。ただし、スキルの低い攻撃者にも攻撃力の高いマルウェアを作成できてしまうリス

---

76 https://www cyberhaven com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/

77 https://openai com/blog/march-20-chatgpt-outage

78 https://www theguardian com/technology/2023/mar/08/darktrace-warns-of-rise-in-ai-enhanced-scams-since-chatgpt-release

79 https://www ncsc gov uk/blog-post/chatgpt-and-large-language-models-whats-the-risk

クは現時点では低いとされている。

2.3.5. 情報環境へのインパクト

＜誤情報の拡散＞

■ Lin ら（2022）[80]は大規模言語モデル（LLM）により生成されたアウトプットの「真実性」（truthfulness）を評価したところ、一般的な誤解を模倣し、人間を欺く可能性のある多くの誤った回答を生成していることを明らかにした。また、最大規模の LLM の真実性が最も低いとし、誤った解答の規模は微妙な不正確さから重大な誤りや「幻覚」まで様々であるとした。これらの結果を踏まえ、誤用や悪用が生じてしまうリスクを指摘した。

■ Google と Microsoft のチャットボット同士がすでにお互いを引用し合うという状況が生まれていることが指摘されている[81]。このことは、生成 AI が生成したコンテンツが、生成 AI のトレーニングデータとして利用されることで、誤情報や偽情報が固定化されてしまう可能性が高まることを示している。AI の言語モデルには、事実と虚構を確実に区別する能力がないため、情報量が増えるとそれだけ学習が強化されていくことになる。

＜悪用・誤用＞

■ 悪意のある行為者がプロパガンダを行うためのコストを大きく下げてしまうことが指摘されている[82]。GPT-3 が公表された際にも、OpenAI 社はこのソフトウェアがスパムやフェイクニュース、プロパガンダの大量作成に利用されることを懸念し、アクセスを制限した。

■ OpenAI社の研究者は、ジョージタウン大学の Center for Security and Emerging Technology および Stanford Internet Observatory と共同で、大規模言語モデル（LLM）が情報操作のためにどのように悪用される可能性があるか、またそうした脅威を軽減させる方法について1年以上にわたって調査し、2023 年 1 月、報告書「生成言語モデルと自動化された影響力行使：新たな脅威と緩和策の可能性」を公表した[83]。影響力行使は ABC[84]、つまり行為者（Actor）、行動（Behavior）、内容（Content）から分析され、軽減策は（1）モデルの構築、（2）モデ

---

[80] Lin, S , Hilton, J & Evans, O TruthfulQA: Measuring How Models Mimic Human Falsehoods https://doi org/10 48550/arXiv 2109 07958 (2022)

[81] https://www theverge com/2023/3/22/23651564/google-microsoft-bard-bing-chatbots-misinformation

[82] https://www oreilly com/radar/ai-powered-misinformation-and-manipulation-at-scale-gpt-3/

[83] https://openai com/blog/forecasting-misuse/

[84] François C Actors, Behaviors, Content: A Disinformation ABC 2019
https://www ivir nl/publicaties/download/ABC_Framework_2019_Sept_2019 pdf

ルへのアクセス、（3）コンテンツの普及、（4）信念の形成、の 4 段階に沿って、技術的実行可能性、社会的実行可能性、ネガティブリスク、効果の 4 側面について検討された。

■ ハーバード大学の Sanders と Scheneier は、ChatGPT がロビイングで活用された場合の民主主義へのリスクを懸念している[85]。

■ OpenAI 社は ChatGPT のコンテンツモデレーションにより、ヘイトスピーチや暴力的なコンテンツ、偽情報や違法行為に関する情報のアウトプットに制限をかけている一方で、それらの制限を回避する方法が開発されていることを Guardian 紙が報告した[86]。

＜アウトプットのアカウンタビリティ＞

■ 大規模言語モデル（LLM）が不正確な情報を生成した場合に誰がその責任を負うことになるのか、明らかにしておく必要がある。そのため、OpenAI 社は利用規約において、個別の法的アドバイス、金融アドバイス、医療アドバイス、そして法執行や刑事司法、移民と亡命に関することなどの高リスクな政府意思決定に利用することを禁じている。

■ Mokander ら（2023）[87]は、既存の AI 監査手続きではうまくいかない原因となる、大規模言語モデル（LLM）特有の 4 つの課題を指摘したうえで、3 つのレベルで監査を行うことを提案している。4 つの課題とは、①生成性（generativity）、②創発性（emergence）、③現実世界モデルの欠如（lack of grounding）、④モデル自体へのアクセスの欠如（lack of access）である。3 つのレベルでの監査とは、①ガバナンス監査、②モデル監査、③適用方法の監査、である。これら 3 つのレベルでの監査結果は相互に監査のインプットとなる。

■ あるテキストが AI によって生成されたものかどうかを検出するためのツールの開発が進められ、生成 AI の開発者と検出ツールの開発者の間の軍拡競争（arms race）に発展すると予測されている。生成 AI によって生成されたテキストには、AI モデル特有の「指紋（fingerprint）」が存在すると考えられている。各モデルの開発には、元となった学習データが異なるため、それぞれのモデルのアウトプットには、学習データの偏りで生じる表現や語彙の特徴や差異が存在する。今後は、これらの指紋を検出することが事業化されることが予測されている一方

[85] https://www nytimes com/2023/01/15/opinion/ai-chatgpt-lobbying-democracy html

[86] https://www theguardian com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards

[87] Mökander J, Schuett J, Kirk HR, Floridi L Auditing large language models: a three-layered approach arXiv 2023 doi: https://doi org/10 48550/arXiv 2302 08500

で、生成 AI のモデルも、指紋の検出を回避するように開発されると考えられている[88]。

2.3.6. 自然環境へのインパクト

- 大規模言語モデル（LLM）の活用の範囲とその規模を考慮すると、優先順位の最も高い検討事項は環境へのインパクトであると Bender ら（2021）[89]が指摘している。なぜならば、LLM の環境コストと経済的コストは、それらの活用からの利益を得る可能性が最も低く、害を受ける可能性が最も高い、疎外されたコミュニティを二重に罰するためである。

- Microsoft 社の Bing 検索エンジンに生成 AI が導入された場合には、従来の検索との比較で、4 倍から 5 倍程度の追加のコンピューティング・パワーを要すると推測されており、ChatGPT が 2021 年以降の情報に対応していないのも、コンピューティングを削減するためであることが報告されている[90]。

- GPT-3 は few-shot generalization という手法を活用するため、各タスクの実施のために度々学習させる必要が軽減され、エネルギー効率は他モデルより優れているが、学習による CO2 の排出は 552 t であり、エネルギー消費は 1287 MWh であることを Patterson ら（2021）[91]が報告している。

2.3.7. 高度な AI システムで生じるリスク

- OpenAI 社は、競争の観点とともに安全性の観点を挙げて GPT-4 の詳細を公開しないと決定したとされるが、LLM のような社会に対するインパクトの大きな技術の中身が一民間企業に委ねられてしまうことの是非は今後議論になると思われる。

- LLM の仕組みがきちんと説明され、人々に理解されないと、米国上院の Murphy 議員（コネチカット州選出）のように、ChatGPT が「高度な化学（chemistry）を独学で学んだ」と主張し、まるで AI が人間のクリエイターから何のインプットもなく、自律的な学んでいるかのように誤解してしまうことが起こりうる[92]。ただし当議員は、AI 研究者らからの指摘を受けて発

---

88 https://www oreilly com/radar/ai-powered-misinformation-and-manipulation-at-scale-gpt-3/

89 https://www theguardian com/technology/2023/mar/08/darktrace-warns-of-rise-in-ai-enhanced-scams-since-chatgpt-release

90 https://www wired com/story/the-generative-ai-search-race-has-a-dirty-secret/?redirectURL=https%3A%2F%2Fwww wired com%2Fstory%2Fthe-generative-ai-search-race-has-a-dirty-secret%2F

91 Patterson D, Gonzalez J, Le Q, Liang C, Munguia LM, Rothchild D, et al Carbon Emissions and Large Neural Network Training arXiv 2021 doi: https://doi org/10 48550/arXiv 2104 10350

92 https://www gizmodo com au/2023/03/how-a-senators-misguided-tweet-can-help-us-understand-ai/

言を修正したとされる。

3. 分野ごとの反応

3.1. 教育分野

- The Daily がソーシャルメディアアプリ Fizz でスタンフォード大学の学生に対して 2022 年 1 月 9 日から 1 月 15 日まで実施した匿名投票によると、4,497 人の回答者の 17%が秋学期の課題や試験を支援するために ChatGPT を使用したと回答したという[93]。約 6 割の学生は「ブレインストーミング、アウトライン作り、アイデア形成」に利用したと回答し、ChatGPT から直接ほとんど編集せずに文章を提出したと回答した学生は 5.5%に過ぎなかった。

- ChatGPT は、教師にとって、テストの採点、報告書やメールの返事を書くといった仕事にかかる時間を減らし、子どもたちに教える時間を増やしてくれるという見方もある[94]。

- Pickell and Doak (2023) [95]は大学教員向けに、レポート作成の剽窃対策として、ChatGPT が対応していない 2022 年以降の観点を求めることや、個人の経験について述べてもらうことなどの対応策を提案している。

- 米ヴァンダービルト大学では、ミシガン州立大学において発生した銃乱射事件を受けて学生に送信したメールに ChatGPT を使用していたことが発覚した。メールには、「Paraphrase from OpenAI's ChatGPT AI language model, personal communication, February 15, 2023」との記載があり、批判が集まった。大学側は、共感を欠く対応について謝罪した[96]。

3.2. マーケティング分野

- Jasper 社はマーケティングに特化した GPT-4 で、ブログ、ソーシャルメディア投稿、ウェブコピー、セールスメール、広告、それらに付随する画像等、顧客向けコンテンツを作成するこ

---

93 https://stanforddaily com/2023/01/22/scores-of-stanford-students-used-chatgpt-on-final-exams-survey-suggests/

94 https://www businessinsider jp/post-264123

95 Pickell, Travis Ryan and Doak, Brian R Five Ideas for How Professors Can Deal with GPT-3 For Now Faculty Publications - George Fox School of Theology 2023; 432 https://digitalcommons georgefox edu/ccs/432/

96 https://www theguardian com/us-news/2023/feb/22/vanderbilt-chatgpt-ai-michigan-shooting-email

とができるとしている[97]。

- DALL-E 2 をはじめとする画像生成ツールは、すでに広告に利用されており、「AI を活用して制作・創作をスムーズに行うという業務プロセスが一般化するでしょう」と指摘されている[98]。

- Noy and Zhang (2023)[90]の公開したプレプリントによると、マーケターや人事担当者、コンサルタント 444 名を対象とした介入研究において、プレスリリース等を含む文章の作成に ChatGPT を活用することで作業時間が 10 分程度短縮することを報告している。同報告は、仕事に対する従事者の満足感を向上させることについても示唆している。

- Salesforce 社は、2023 年 3 月、CRM（Customer Relationship Management）向けの生成 AI である Einstein GPT を発表した。Einstein GPT を活用することで、マーケティングの観点ではパーソナライズされたコンテンツを動的に生成し、電子メール、モバイル、Web、広告を通して顧客や見込み客を獲得することができるとしている[99]。

3.3. 学術出版分野

- Nature 誌は、すべての Springer Nature ジャーナルとともに、2 つの原則を策定し、既存の著者向けガイド[100]に追加した[101]。1 点目は、大規模言語モデル（LLM）ツールが研究論文のクレジットされた著者として認められることはないだろうということ。著作権の帰属には研究成果に対する説明責任を伴うが、AI ツールにはそのような責任を負わせることができないからである。2 点目は、LLM ツールを使用する研究者は、その使用方法を「方法」か「謝辞」に記載する必要があるということ。論文にこれらの項目がない場合は序文か他の適切なセクションに記載すべきである。

- Science 誌は、AI ツールを著者として認めないことを明示している。それだけでなく、AI、機械学習、類似のアルゴリズムツールから生成された文章は編集者の明確な許可なしに使用することができないとしている。加えて、図や画像、グラフィックについてもこれらのツール

---

97 https://www jasper ai/

98 https://dentsu-ho com/articles/8322

99 https://www salesforce com/news/press-releases/2023/03/07/einstein-generative-ai/

100 https://www nature com/nature/for-authors/initial-submission

101 Editorial Tools such as ChatGPT threaten transparent science; here are our ground rules for their use, Nature 2023; 613: 612 doi: https://doi org/10 1038/d41586-023-00191-1

から生成されたものは認められない。なお、このポリシーに反する場合は研究不正に該当するとの記載もある[102,103]。

- PNAS 誌は、PNAS 誌および PNAS Nexus 誌のオーサーシップおよび編集方針を更新し、方法（これがない場合は、謝辞）に生成 AI を使用したことを記載する必要があるとした。また、オーサーシップの基準は満たさないものとしている。これは、透明性と説明責任を担保するためのものである[104]。

- Elsevier 社は、著者が執筆プロセスにおいて生成 AI や AI 支援技術を使用する場合、これらの技術は読みやすさや言語を改善するためにのみ使用されるべきであるとしている。AI は権威的に聞こえる出力を生成することができるが、不正確であったり、不完全であったり、偏っていたりすることがあるため、技術の適用は人間の監視と管理の下で行われるべきであり、著者は結果を慎重に確認し編集する必要がある。著作物の内容については、著者が最終的な責任と説明責任を負う。著者は、AI および AI 支援技術の使用について原稿で開示する必要があり、公表された著作物にはステートメントが掲載される予定である。これらの技術の使用を宣言することは、著者、読者、査読者、編集者、投稿者間の透明性と信頼をサポートするものであり、関連するツールや技術の使用条件の遵守を容易にする[105]。

- Taylor & Francis 社は、学術研究において AI ツールの利用が増加していることを認識しており、適切かつ責任を持って使用される場合、そのようなツールは研究成果を増大させ、知識による進歩を促進する可能性を秘めていると考えていとの見解を示した。その上で、著者としては認められず、使用について適切に記載されなければならないとしている[106]。

- プレプリントサーバ arXiv は、生成 AI 言語ツールが有用で役立つ結果を生み出すだけでなく、エラーや誤解を招く結果を生み出す可能性があることに注目している。このため、使用について著者に方法として適切に記載するように求めると共に不適切な言語、剽窃されたコンテンツ、偏ったコンテンツ、エラー、間違い、誤った参照、または誤解を招くコンテンツを生成し、その出力が科学的著作物に含まれる場合、それは著者の責任であることを述べている[107]。

---

102 https://www science org/content/page/science-journals-editorial-policies

103 井出和希　生成 AI とオーサーシップ：国際誌の対応動向　日本薬理学雑誌 2023, in press

104 https://www pnas org/post/update/pnas-policy-for-chatgpt-generative-ai

105 https://www elsevier com/about/policies/publishing-ethics#Authors

106 https://newsroom taylorandfrancisgroup com/taylor-francis-clarifies-the-responsible-use-of-ai-tools-in-academic-content-creation/

107 https://info arxiv org/help/moderation/index html#policy-for-authors-use-of-generative-ai-language-tools

■ これまで、AI は学術界において盗作を防ぐためのツールであったが、Dehouche（2021）[108]は、生成 AI の開発により、従来の「盗用（plagiarism）」の定義を再考する必要があると主張している。

3.4. ジャーナリズム分野

■ 2023 年 1 月 12 日付の Futurism の記事で、2022 年 11 月頃から CNET が、金融セクションの解説記事に"CNET Money Staff "という呼称で「自動化技術」を使っていた、つまり AI が自動で記事を作成していたことが明らかにされた[109]。AI による記事作成についての公式発表はなく、"CNET Money Staff"をクリックして初めて「この記事は、AI エンジンによる支援と、編集スタッフによる審査、事実確認、編集が行われた。」[110]と表示されるため、多くの読者はそのことに気づかなかった。Red Venture 社が所有する同社はその後、AI ツールの使用を一時停止するとともに、この技術の助けを借りて書かれた記事の半分以上に間違いが見つかったという[111]。

■ デジタルメディア企業の BuzzFeed の CEO である Jonath Peretti 氏は 1 月末、スタッフへのメモにおいて、個人向けにカスタマイズされたコンテンツなどを OpenAI 社の技術を使って作成するなど、運営に AI を導入する意向を示したことが Wall Street Journal 紙によって報じられた[112]。これを受けて株価は 2 倍以上に跳ね上がったという。

■ ドイツのメディアグループ Axel Springer 社は、自動化と生成 AI の導入により、人員削減の準備を進め、今後は調査報道と独自のコメンタリーに注力することを発表した。英国の新聞紙 Daily Mirror と Daily Express においても、ChatGPT の可能性を探る検討会が立ち上げられたという[113]。

■ 出版業界は、ChatGPT などの AI ツールのトレーニングに自社コンテンツがどれくらい使われているかに関心をもっているという[114]。調査結果によっては補償を求めたり、法的措置を

---

[108] Dehouche N Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3) Ethics in Science and Environmental Politics 2021; 21: 17-23 doi: https://doi org/10 3354/esep00195

[109] https://futurism com/the-byte/cnet-publishing-articles-by-ai

[110] https://www cnet com/profiles/cnet%20money/

[111] https://www theverge com/2023/1/26/23572834/buzzfeed-using-ai-tools-personalize-generate-content-openai

[112] https://www wsj com/articles/buzzfeed-to-use-chatgpt-creator-openai-to-help-create-some-of-its-content-11674752660?mod=djemalertNEWS

[113] https://www theguardian com/technology/2023/mar/01/german-publisher-axel-springer-says-journalists-could-be-replaced-by-ai

[114] https://www wsj com/articles/publishers-prepare-for-showdown-with-microsoft-google-over-ai-tools-6514a49e

行ったりする可能性も示唆されている。生成 AI ツールの回答においてニュースソースへのリンクが提供されるかどうか、そのリンクへのクリック率がどれくらいかにも依存する。テック業界と出版業界の間ではこれまでもコンテンツの使用料について論争が続いてきたが、生成 AI のトレーニングデータとしての使用という新たな問題が追加された形となる。

■ オックスフォード大学にあるロイタージャーナリズム研究所は、AI がジャーナリズムに与える影響について、5 人の AI 専門家の意見を紹介した[115]。彼らは、ChatGPT のような AI 技術が、情報収集や報道の自動化などの分野での利用に大きな潜在的な価値があるうえに、AI が報道の偏りや不正確な情報を排除することができる可能性もあると指摘した。他方、彼らはまた、AI が信頼性のある情報源との区別をつけることや倫理的な課題に対処することが困難であることも指摘した。結論としては、AI が人間のジャーナリストを完全に置き換えることはないだろうと考えている。

3.5. エンターテインメント分野

■ 2023 年 1 月 31 日、Netflix は AI 生成画像を背景に採用したアニメ「犬と少年」を公開した[116]。アニメ制作と並行してアニメ背景画生成ツールの共同開発にも挑戦し、アニメの背景美術制作を最新技術によって補助できるかどうかの実験と位置付けられている。背景デザイナーには「AI（+Human）」と記載されている。

■ 2022 年 10 月 20 日、イラスト・漫画・小説の投稿や閲覧を行うインターネットサービスである pixiv は、AI 生成作品の取り扱いに関するサービスの方針を出した[117]。その後、2022 年 10 月 30 日、AI 生成作品の取り扱いに関する機能をリリースした[118]。投稿時に使用の有無を記載したり、検索時にフィルタリングにより表示を減らしたりすることができるとされている。また、ランキングにおいても AI 生成作品を他の作品と分けるとしている。

3.6. 司法分野

■ コロンビアのカルタヘナ市の第 1 巡回裁判所を管轄する Juan Manuel Padilla Garcia 判事は、

115 https://reutersinstitute politics ox ac uk/news/chatgpt-threat-or-opportunity-journalism-five-ai-experts-weigh

116 https://about netflix com/ja/news/the-dog-and-the-boy

117 https://www pixiv net/info php?id=8710

118 https://www pixiv net/info php?id=8728&lang=ja

ChatGPT を使って事件に関する法的質問を投げかけ、ChatGPT とのやりとりが 2023 年 1 月 30 日付の裁判文書に記載されている[119]。ただし、AI は主に判決の起草を効率的に進めるために使用され、その回答はきちんと事実確認されたことが書かれている。

- OpenAI 社の GPT-4 は、米国の司法試験の模擬試験において、受験者の上位 10%程度のスコアで合格したとされる[120]。

- 弁護士の松尾剛行氏は ChatGPT 等の大規模言語モデル（LLM）の弁護士実務への影響をまとめた文書において、技術的制約として、1）根拠が分からない（不透明）なこと、2）新しいこと/データが少ないことに答えられないこと、3）本質的には「分かっていない」まま「データが多い」分野について振る舞いが上手くなっていくだけであること、4）操作・攻撃の可能性、5）責任を取らないこと、6）コミュニケーション、を挙げた[121]。そのうえで、弁護士自身ができることの支援に使ったり、弁護士自身ができないことを実施したり、それらのことを通して業務を可視化するといった活用の方向性を示唆した。

3.7. 医療分野

- GPT-3 の医療用の活用は OpenAI 社によってサポートされていないが、複数の研究者及び企業によって活用が検討されていると Quanch（2022）[122]が報告する。フランスの Nabla 社が実験用に、GPT-3 を活用した医療用チャットボットを開発し、患者との模擬セッションの中で、不適切なアドバイスを提供したと報告した。その中では、「とても調子が悪いので、自殺した方が良いですか？」と尋ねた模擬患者に対して、チャットボットは「そうすべきだと思います。」と反応したと報告されている。

- ChatGPT は、退院レポートの作成プロセスの最初のステップとして活用され、その後、医師がアウトプットを確認することを Patel と Lam（2023）[123]が期待している。将来の課題は、この技術を採用するかどうかではなく、どのように採用するか、であると主張している。

---

[119] https://www vice com/en/article/k7bdmv/judge-used-chatgpt-to-make-court-decision

[120] https://openai com/research/gpt-4

[121] https://note com/matsuo1984/n/n006e3e569eb0

[122] https://www theregister com/2020/10/28/gpt3_medical_chatbot_experiment/

[123] Patel SB, Lam K ChatGPT: the future of discharge summaries? The Lancet Digital Health 2023; 5: e107-e108 doi: https://doi org/10 1016/S2589-7500(23)00021-3

しかし、Lancet Digital Health の論説[124]では、Patel と Lam の論文に含まれた、ChatGPT によって生成された退院サマリーに誤りがあったことを指摘し、プロンプトに含まれていなかった情報がサマリーに追加されたことを指摘している。生成 AI の活用により、学術的出版物には、誤りだけではなく、盗用されたものが組み込まれるリスクが生じ得ることから、科学的記録の完全性に深刻な影響を及ぼし得る。また、誤った情報に基づいた研究や政策の決定が進められる可能性もあることを懸念している。

- Mbakwe ら（2023）[125]は、ChatGPT が米国医療免許試験（USMLE）に合格同等のパフォーマンスを示したことを受け、医学教育と USMLE に関する懸念を述べている。特に、潜在的なバイアスが含まれていながらも、ChatGPT の学習データとなった、インターネットの医療コンテンツがこの試験に合格するために十分であったことを問題視している。また、あらゆる情報にすぐにアクセスできる今の時代では、USMLE に合格することは、果たして医師の能力を示すのに十分な基準であるかどうかを問う。

- Sarraju ら（2023）[126]は、心血管疾患の予防アドバイスに ChatGPT を用いる試行を行い、25 項目のうち 21 項目（84%）で適切に回答したとされている。また、3 回の応答の一貫性をもとにして「信頼できない」と判断された回答はなく、「単純な」質問への回答について、インタラクティブに活用していくことができる可能性を提示した。しかしながら、同様に単純な質問であってもあらゆる医療分野で適切な回答が返ってくる訳ではないことには留意を要する。

- Haupt ら（2023）[127]は、ヘルスケアでは、GPT は研究、教育、臨床ケアにおいて役割を果たすことができると述べている。研究の場では、科学者が質問を立て、研究プロトコルを作成し、データを要約するのに役立つ。医学教育では、GPT はインタラクティブな百科事典の役割を果たすことができる。また、患者とのやり取りをシミュレートすることで、病歴聴取のスキルを磨くこともできる。また、学生が授業や病棟で作成する記録やケアプランなどの草稿を作成することも可能としている。また、臨床医にとっては、GPT が反復作業を担うことで、燃え尽き症候群を緩和できる可能性がある。加えて、臨床的な意思決定をサポートし、電子

---

[124] The Lancet Digital Health ChatGPT: friend or foe? The Lancet Digital Health 2023; 5: e102 doi: https://doi org/10 1016/S2589-7500(23)00023-7

[125] Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A ChatGPT passing USMLE shines a spotlight on the flaws of medical education PLOS Digital Health 2023; 2: e0000205 doi: https://doi org/10 1371/journal pdig 0000205

[126] Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model JAMA 2023; 329: 842-844 doi: https://doi org/10 1001/jama 2023 1044

[127] Haupt CE, Marks M AI-Generated Medical Advice—GPT and Beyond JAMA 2023, in press doi: https://doi org/10 1001/jama 2023 5321

カルテプラットフォームにも組み込むことができるだろう。そして、UpToDate のような頻繁に使用するリソースを補強したり、置き換えたりするかもしれない。理論的には、医師はこのソフトウェアに患者情報を入力し、診断や予備的な治療計画についての情報を求めることもできるが、現在のところ医療保険の相互運用性と説明責任に関する法律（Health Insurance Portability and Accountability Act, HIPAA）に準拠しておらず、患者のプライバシーを危険にさらす可能性があることも指摘している。

4. おわりに：ELSI への対応動向

- EU では、文章や画像を生成する生成 AI モデルは「汎用目的型 AI（general-purpose AI）」システムと呼ばれるカテゴリーに含まれることになる。2021 年 4 月に欧州委員会が提案した AI 規制法案は、2023 年 3 月段階では欧州議会で議論が続いているが、規制対象となる AI の定義については政治的な合意に達し、OECD において利用されている定義とほぼ同様なものとすることになったと報道された[128]。定義の中での「予測」にはコンテンツの予測も含まれるとすることで、ChatGPT のような生成 AI モデルも対象となりうるような措置がとられているという。ただし、高リスク AI システムに該当するか、汎用目的型 AI としてそれらの要件が免除されるのかについてはまだ定まっていない。

- 米国の著作権局（Copyright Office）は 3 月 16 日、生成 AI 技術の急速な発展を受けて、AI ツールを用いて生成された作品の著作権の範囲や、AI のトレーニングにおける著作物の使用など、AI が提起する著作権法および政策上の問題を検討する新たなイニシアチブを開始した[129]。同時に、著作権の登録においては、申請者が作品に AI で生成されたコンテンツが含まれていることを開示する義務があることを明確にする指針を官報に公表した[130]。

- 非営利団体である Future of Life Institute は 3 月 29 日、「すべての AI 研究機関に対し、GPT-4 よりも強力な AI システムの訓練を少なくとも 6 ヶ月間、直ちに一時停止するよう要請する」とするオープンレターを公表し、多くの著名人が署名に名を連ねた[131]。最近の数カ月の発展は、その作成者さえも理解できたり、予測できたり、確実に制御したりすることが困難である

[128] https://www euractiv com/section/artificial-intelligence/news/eu-lawmakers-set-to-settle-on-oecd-definition-for-artificial-intelligence/

[129] https://www copyright gov/newsnet/2023/1004 html

[130] https://www govinfo gov/content/pkg/FR-2023-03-16/pdf/2023-05321 pdf

[131] https://futureoflife org/open-letter/pause-giant-ai-experiments/

ことが明らかであり、人間が制御不能な競争状態に陥っていると現状を分析したうえで、「強力な AI システムは、その効果が肯定的であり、リスクが管理可能であると確信した場合にのみ開発されるべき」とした。また、この休止期間の間に、独立した外部の専門家と協力して、共有された安全プロトコルを開発・実装することを提案した。そしてこれと並行して、AI 開発者は政策立案者と協力して、強固な AI ガバナンスシステムの導入を劇的に加速させなければならないとした。

元 Google の Gebru らが設立した DAIR 研究所のメンバーはオープンレターに対して 3 月 31 日、レターに引用された「確率的オウム」論文の著者として声明を出した[132]。レターでは生成 AI について恐怖心と AI ハイプを煽るものになっており、このような仮想的なリスクは、「長期主義（longtermism）」と呼ばれる危険なイデオロギーの中心的な考え方であり、今日の AI システムの展開によって生じる実際の害を無視するものだとした。そして、必要なものは透明性を強制する規制であると主張した。

プリンストン大学の Kapoor と Narayanan もオープンレターを、現実のリスクを無視した SF（サイエンスフィクション）の危険についてのミスリーディングなものであるとしてオープンレターを批判している[133]。誤報、労働への影響、安全性の 3 つが AI の主要なリスクであることに同意するものの、オープンレターが取り上げるそれら内容は空想上の未来のリスクであるとして、次のような表を掲載している（表 1）。

表 1：Kapoor と Narayanan が指摘する AI のリスク

	空想上のリスク（Speculative risks）	現実のリスク（Real risks）
誤報（misinformation）	悪意ある誤報	不正確なツールへの過度の依存
労働への影響	LLM がすべての仕事に取って代わる	中央集権化した権力、労働の搾取
安全性	長期の人類存続リスク	直近のセキュリティリスク

■ 米国の非営利研究機関である「AI・デジタル政策センター（Center for AI and Digital policy：

[132] https://www dair-institute org/blog/letter-statement-March2023

[133] https://aisnakeoil substack com/p/a-misleading-open-letter-about-sci

CAIDP）」は 3 月 30 日、OpenAI 社による生成 AI 実験が「偏った、欺瞞的な、プライバシーと公共の安全に対するリスク」であり、これは連邦取引委員会（FTC）が AI の利用に求める「透明で、説明可能で、公平で、アカウンタビリティを果たしながら経験的に確かである」という条件を一つも満たしていないとして、FTC に対して OpenAI 社への調査を開始し、さらなる商業的なリリースを差し止めることを要請した[134]。具体的には、バイアス、子どもの安全、消費者保護、サイバーセキュリティ、欺瞞、プライバシー、透明性、公共の安全などが懸念事項として挙げられた。

■ イタリア政府のデータ保護機関（GPDP）は 3 月 31 日、欧州の一般データ保護規則（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ただ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[135]。GDPR 違反として挙げられたのは以下の 3 点である。

a. ChatGPT のユーザーの会話やサービス加入者の支払い情報に影響を与えるデータ侵害が 3 月 20 日に報告されたこと、

b. 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、

c. OpenAI 社の利用規約では 13 歳を超えるユーザーを対象としているにもかかわらず年齢確認の仕組みが一切ないこと

そして、この命令を遵守するために実施した措置について、20 日以内にイタリアのデータ保護機関に通知する必要があり、そうでない場合は、2000 万ユーロまたは全世界の年間総売上高の 4%を上限とする罰金が課される可能性がある。

実際、イタリアでの ChatGPT のサービス提供は止められた。

■ 英国政府は AI に対する監督について、規制的アプローチを推進する EU と異なり、柔軟でイノベーション促進型のアプローチを支持している[136]。政府最高科学顧問および国家技術顧問であるパトリック ヴァランス卿による新規技術の「イノベーション促進型の（Pro-innovation）」

[134] https://www caidp org/cases/openai/

[135] https://www gpdp it/web/guest/home/docweb/-/docweb-display/docweb/9870847 本文（イタリア語）は、https://www garanteprivacy it/web/guest/home/docweb/-/docweb-display/docweb/9870832

[136] https://www gov uk/government/publications/ai-regulation-a-pro-innovation-approach

規制に関するレビューが実施され、サンドボックスやテストベッドなどを通して実験を促進し、技術が確立された段階で規制の国際的な調和を図り、イノベーティブな企業の参入を確保するというアプローチが提案された[137]。その中で、生成 AI も対象となり、「政府は、知的財産法と生成 AI の関係について明確な政策的見解を公表し、イノベーターや投資家に確信を提供すべきである。」という提言を行った。現在、知的財産法と生成 AI の関係が不明確であることがステークホルダーから指摘されていることから、データ、テキスト、および画像のマイニングを可能とする方向で規制内容を明確にすることは、英国の AI 産業とクリエイティブ産業の成功の基礎となることが指摘された。これに対して英国政府は、知的財産庁（IPO）が、AI 企業がモデルへの入力として著作物にアクセスすることを支援する一方で、著作物の権利者を支援するために生成された出力が保護されることを保証する指針を提供する行動規範を夏までに作成する予定であると回答した。

[137] https://www gov uk/government/publications/pro-innovation-regulation-of-technologies-review-digital-technologies

ELSI NOTE No. 26

生成 AI（Generative AI）の倫理的・法的・社会的課題（ELSI）論点の概観 2023 年 3 月版

令和 5 年 4 月 10 日

大阪大学 社会技術共創研究センター
Research Center on Ethical, Legal and Social Issues

〒565-0871 大阪府吹田市山田丘 2-8
大阪大学吹田キャンパステクノアライアンス C 棟 6 階
TEL 06-6105-6084
https://elsi.osaka-u.ac.jp

大阪大学

Osaka University
Research Center on
Ethical, Legal and
Social Issues