提言

生成 AI を受容・活用する社会の実現に向けて

令和７年（２０２５年）２月２７日

日　本　学　術　会　議

この提言は、日本学術会議情報学委員会が中心となり審議を行ったものであり、日本学術会議として公表するものである。

日本学術会議情報学委員会

委員長	下條　真司	（第三部会員）	青森大学ソフトウェア情報学部教授／大阪大学名誉教授
副委員長	高田　広章	（第三部会員）	名古屋大学未来社会創造機構教授
幹　事	黒橋　禎夫	（第三部会員）	大学共同利用機関法人情報・システム研究機構国立情報学研究所所長／京都大学大学院情報学研究科特定教授
幹　事	佐古　和恵	（第三部会員）	早稲田大学理工学術院教授
	浅川智恵子	（第三部会員）	IBM Fellow／日本科学未来館館長／ Carnegie Mellon University IBM 特別功労教授
	有村　博紀	（第三部会員）	北海道大学大学院情報科学研究院教授
	内田　誠一	（第三部会員）	九州大学理事／副学長
	大場みち子	（第三部会員）	京都橘大学工学部情報工学科教授
	田浦健次朗	（第三部会員）	東京大学執行役／副学長
	永井由佳里	（第三部会員）	北陸先端科学技術大学院大学理事／副学長

本提言の作成にあたり、以下の方々に御協力いただいた。

連携会員	相澤　彰子	大学共同利用機関法人情報・システム研究機構国立情報学研究所副所長／コンテンツ科学研究系教授
連携会員	上田　修功	国立研究開発法人理化学研究所革新知能統合研究センター副センター長／日本電信電話株式会社 NTT コミュニケーション科学基礎研究所客員フェロー
連携会員	佐藤　一郎	大学共同利用機関法人情報・システム研究機構国立情報学研究所情報社会相関研究系教授
	生貝　直人	一橋大学大学院法学研究科教授
	井尻　善久	SB Intuitions 株式会社取締役兼 CRO 兼 R&D 本部長
	越前　　功	大学共同利用機関法人情報・システム研究機構国立情報学研究所情報社会相関研究系教授／シンセティックメディア国際研究センターセンター長
	岡崎　直観	東京科学大学情報理工学院情報工学系教授
	尾形　哲也	早稲田大学理工学術院教授
	岡野原大輔	株式会社 Preferred Networks 代表取締役最高研究責任者
	河原　大輔	早稲田大学理工学術院教授
	佐藤　　健	大学共同利用機関法人情報・システム研究機構人工知能法学研究支援センターセンター長／特任教授
	佐藤　真一	大学共同利用機関法人情報・システム研究機構国立情報学研究所コンテンツ科学研究系教授／主幹
	杉山　弘晃	日本電信電話株式会社 NTT コミュニケーション科学基礎研究所主任研究員
	鈴木　　潤	東北大学言語 AI 研究センターセンター長／教授

関根　聡	国立研究開発法人理化学研究所革新知能統合研究センターチームリーダー／大学共同利用機関法人情報・システム研究機構国立情報学研究所大規模言語モデル研究開発センター特任教授
武田　浩一	大学共同利用機関法人情報・システム研究機構国立情報学研究所大規模言語モデル研究開発センター副センター長／特任教授
西貝　吉晃	千葉大学大学院社会科学研究院教授
羽深　宏樹	京都大学大学院法学研究科特任教授／スマートガバナンス株式会社代表取締役 CEO
福島　俊一	国立研究開発法人科学技術振興機構研究開発戦略センターフェロー
前田　健	神戸大学大学院法学研究科教授
宮尾　祐介	東京大学大学院情報理工学系研究科教授
山岸　順一	大学共同利用機関法人情報・システム研究機構国立情報学研究所コンテンツ科学研究系教授／総合研究大学院大学先端学術院教授
横田　理央	東京科学大学総合研究院スーパーコンピューティング研究センター教授

本提言の作成にあたり、以下の職員が事務および調査を担当した。

事務	新田　浩史	参事官（審議第二担当）（令和６年８月から）
	角田美知子	参事官（審議第二担当）付参事官補佐（令和６年 10 月から）
	藤田　崇志	参事官（審議第二担当）付審議専門職
調査	辻　政俊	上席学術調査員

要　　旨

１　作成の背景

急速に進展する生成 AI には、あらゆる学術分野、産業分野、そして社会全体に大きな影響を持つという包括性、将来的には人間と共存する知的レベルとなり得る革新性、さらに、それが加速度的に進展するという加速性などの特徴がある。それゆえに、脅威や課題が存在するとともに、社会への大きな波及効果があり、人類社会の重要課題に対して解決策を提供する可能性がある。

このような状況のもとで、生成 AI の現状と動向、脅威と課題、活用による波及効果について、学術の立場から深く洞察し、生成 AI を受容・活用する社会の実現に向けてどのような施策をとるべきかについて提言をまとめる。

２　現状および問題点

生成 AI は、2020 年代以降急速にその技術が発展・普及し、特に大規模言語モデル（Large Language Model：LLM）を基盤とする ChatGPT は、2022 年 11 月に公開後２ヶ月でアクティブユーザー数が１億人に達した。LLM は Transformer と呼ばれるニューラルネットワークモデルが基盤となっており、モデルサイズ（パラメータ数）と学習データ量に対して、対数スケールで性能が向上することが知られている。言語だけでなく、画像・映像、音声・音楽などと統合するマルチモーダル処理、ロボティクスへの応用も進んでいる。

生成 AI には様々なリスクが存在する。生成 AI の活用においては、ハルシネーション（事実と異なる情報を出力すること）、高品質な生成メディアによる詐欺や世論操作、非社会的な回答の生成、機密情報漏洩などの懸念がある。また、著作権侵害、名誉毀損、芸術的活動への脅威、社会の価値観や文化への影響などの懸念もある。これらの問題に対処するために、生成 AI モデルには正確性、指示追従性、頑健性、透明性、説明可能性などが求められる。

一方で、生成 AI には人類社会に資するプラスの面が大いに期待されている。科学技術においては、生成 AI の活用により仮説生成、仮説検証、論文による知識流通など科学技術の各ステップが大きく進展しようとしている。知識の細分化・専門化による学問領域の分断が指摘される中、生成 AI の活用による分野横断的な新たな知の創造も期待される。産業界においても、生成 AI の活用による業務効率化・業務量削減が、労働力不足や長時間労働などの社会課題を解決する強力な一手となり得る。さらには、教育の効率化や高度化、文章の推敲・要約・翻訳、挨拶文やメールの作成支援、旅行計画の提案、投資のアドバイス、音楽・アート・デザインの生成など、我々の日常的な活動への波及も始まっている。

３　提言の内容

生成 AI の世界的進展が留まる気配のない中で、我が国は、リスク対策についても十分に工夫をしながら、生成 AI の研究開発や社会での活用を積極的に進め、人類と AI の共存社会のデザインで世界をリードすべきである。

⑴　生成 AI 研究開発の望ましい体制

① 日本の技術競争力を強化するため、国家戦略として生成 AI の研究開発を推進すべきである。特に、オープンな研究開発の取り組みへの支援を重視・強化することが必要である。

② 日本国内の生成 AI 研究者コミュニティの強化と国際的研究連携の推進が必要である。プライバシーやセキュリティに配慮したデータインフラの構築を支援するとともに、公共データの開放や産業界とのデータ共有プロジェクトを奨励すべきである。

③ 生成 AI による判断や行動が、人間の価値観や倫理観に合うことが極めて重要であり、学習データや学習手法を含む開発プロセスの透明性を確保すること、AI の設計・開発・評価においてガイドラインを作成してリスクを最小化すること、AI ガバナンスの国際的ルールメイキングに日本の考え方を反映させる体制を構築することが必要である。

⑵　生成 AI モデルの適切な運用

① 生成 AI モデルがサイバー攻撃や物理的攻撃から適切に保護される必要があり、これらの攻撃を検知・回避する頑健なシステムが構築されるべきである。

② AI 技術に起因する問題が発生した場合に、迅速かつ適切に対処できる体制を整えることが必要である。また、国際的な協力を通じて、AI 技術の標準化やベストプラクティスを共有し、グローバルな視点での AI の発展と運用を推進することが重要である。

③ 人間中心の原則に基づく持続可能な社会の実現に向けて、市場原理や競争原理に任せるのではなく、地球規模の課題や社会・経済にとって最重要な問題に対して AI の利活用・運用を優先すべきである。

⑶　責任ある生成 AI 実装に向けた制度設計

① 従来型の規制モデルでは、複雑で変化の速い AI がもたらす様々なリスクに適切に対処することができない。アジャイル（迅速かつ反復的）でマルチステークホルダー型のガバナンスを志向すべきである。

② 政府は、オープンな場でのルール形成の促進、事故調査への関係者の積極的な協力を促す制度設計、事故被害者に対する迅速な救済制度の設計などを行う必要がある。

③ 民間主体は、政府を含むステークホルダーに十分な質と量の情報開示を行うとともに、ステークホルダーからのフィードバックを得て、常にガバナンスのあり方を改善する必要がある。

(4) 生成 AI モデル以降の教育とリテラシー

① 社会全体で生成 AI の教育やリスキリングに取り組んでいくことが必要であり、それを推進するためのリテラシーを持つ人材の養成と教育プログラムの推進、リスキリング支援があまねく必要である。この際、地域格差に配慮し、むしろ地域格差を解消することを目指すべきである。

② 慎重な議論を行った上で、AI の活用を前提として AI との共存を目指した新たな教育への転換を図るべきである。従来の知識の伝達に偏重するのではなく、AI を批判的に利用し、課題を解決し、創造する能力を高める教育・カリキュラムが必要である。また、新たな教育について情報共有・議論する場を国として支援することも重要である。

③ AI の活用は、学術を学際的に深め、複合的な社会課題の解決につながるが、そのためには、科学者が高い AI リテラシーを身につけることが必要であり、学術分野間および産学間の対話・連携の促進が必要である。

目　　次

１　はじめに

生成 AI は、2020 年代以降急速にその技術が発展・普及し、特に大規模言語モデル（Large Language Model：LLM）を基盤とする ChatGPT は、2022 年 11 月に公開後２ヶ月でアクティブユーザー数が１億人に達し、その後、教育・研究を含め社会全体に大きな影響をもたらしている[1]。国内においても、2023 年７月４日に文部科学省から初等中等教育段階における生成 AI の利用について、暫定的なガイドラインが発行され[1]、大学など大多数の高等教育機関でも独自にガイドラインが策定・公開されている。その趣旨は、生成 AI の持つ高いポテンシャルを有効利用することの促進と、プライバシーや著作権の侵害といった利用上のリスクや不正な使用に対する注意喚起の二つの側面を持つ。

利活用の促進とリスクへの対策という両面への配慮は、その後の議論においても主要な論点となっている。2023 年５月に G7 関係閣僚を中心に AI の活用や開発・規制に関する国際的なルール作りを推進するプロセス（広島 AI プロセス）が開始され、12 月の閣僚級会合でまとめられた「広島 AI プロセス包括的政策枠組み」が G7 首脳に承認された[2]。また、2024 年２月に独立行政法人情報処理推進機構に AI セーフティ・インスティテュート（AISI）という AI の安全性評価に関する基準や手法を検討する専門の機関が設立されている。2024 年４月には、日本学術会議も参加した G サイエンス学術会議 2024 において「人工知能と社会」についての共同声明が取りまとめられ公表された[3]。また、2024 年５月には欧州連合において AI 法（The Artificial Intelligence Act：AI Act）が採択され[2]、今後国際的にも大きな影響を持つことが予想される。2024 年７月には、「地球規模の変革に向けた科学」というテーマのもとブラジルで開催されたサイエンス 20 において、サブテーマとして人工知能の倫理や社会的影響が議論され、共同声明が公表されている[4]。

生成 AI に関連する問題を考える難しさとして、技術の進展が極めて早いこと、教育・研究から産業、人々の仕事・暮らしに至るまでその影響する範囲が極めて広いこと、将来的には人類と AI の共存社会の到来が予想されるがその姿を正確に予見することは極めて困難であること、などが挙げられる。

このような状況のもとで、生成 AI を受容・活用する社会の実現に向けて、学術の立場から生成 AI について深く洞察し、前述の二つの側面を調和させたどのような施策をとるべきかについて提言をまとめる。

なお、生成 AI はまさに日進月歩で急速に進展している。本提言は主に 2024 年８月から 10 月にかけて執筆したものであり、その時点の技術的・社会的状況を踏まえたものである。また、一部でロボットなどとの関係に言及しているが、基本的にはサイバー空間での問題を議論している。自動運転、工場の自動化、医療ロボットなどサイバー空間と物理空間を繋ぐ問題は対象外とする。同様に AI の軍事応用の問題なども対象外とする。

---

[1] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[2] https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/?trk=public_post_comment-text

2　生成 AI の現状と動向

(1)　生成 AI と基盤モデル

まず、生成 AI および基盤モデルという用語・概念を歴史的経緯とともに整理し、その上で現在の生成 AI の主流モデルとなっている Transformer について簡単な説明を行う。

「基盤モデル（Foundation Model）」という用語は、2021 年にスタンフォード大学のワーキンググループによって命名され、Bommasani らの“On the Opportunities and Risks of Foundation Models”という論文[5]で初めて世の中に紹介されたとされている。この論文で紹介されている基盤モデルの定義は「大量かつ多様なデータで学習された様々な下流タスクに適応できるモデル」となっている。この定義に従うと、この論文が発表されるかなり前から基盤モデルと呼ぶことができるモデルが存在していたと言える。例えば、言語処理分野で、文章中の穴埋め問題から単語のベクトル表現を学習する word2vec（2013 年）[6]、あるいは、画像処理分野で、大規模画像データベース ImageNet によって画像分類を学習した VGG（2014 年）[7]、ResNet（2015 年）[8]などが初期の基盤モデルに相当する。

一方、「生成 AI（Generative AI）」という用語に関しては、いつ誰が使い始めたかの定説はない。一説によると、敵対的生成ネットワーク（Generative Adversarial Network：GAN）[9]が 2014 年に提案されてから、しだいに生成ネットワークや生成器（Generator）という用語が、生成 AI という使われ方に変化したと言われている。また、「生成モデル（Generative Model）」という用語自体は、確率モデルに基づくアプローチが盛んに取り組まれていた 2000 年以前から用いられていた用語である。また、2018 年に発表された OpenAI の Radford らの論文“Improving Language Understanding by Generative Pre-Training”[10]、いわゆる GPT の最初の論文において Generative Pre-Training という用語が使われている。このように、Generative X のような用語が様々に使われていた状況の中から、生成 AI（Generative AI）という用語が、自然発生的に使われるようになったと想像される。

基盤モデル、生成 AI という用語・概念が生まれた歴史的経緯は上述の通りであるが、基盤モデルの方が必ずしも生成能力があることを要件としないという意味で、より広い概念・用語であると言える。一方で、昨今の社会への大きなインパクトは、生成 AI が言語や画像を生成し、人間との知的な対話やインタラクションが可能となった点に起因する。その意味で本提言のタイトルには、これらの技術の総称として「生成 AI」という用語を採用したが、本提言の説明では文脈に応じて適宜二つの用語を使い分けることとする。

生成 AI について、言語、画像、音声、動画など、何を生成するかによって、そのモデル構成は異なり得る。その中で、文章を生成する生成 AI は、文脈が与えられた時に次の単語を予測する分類問題の連続とみなすことができ、Transformer[11]と呼ばれる可変長の入力系列および出力系列を扱うのに適したニューラルネットワークが主流となっている。ここでの入力系列と出力系列とは、主に入力または出力される単語の列である。さらに、現時点では、Transformer の構成が、様々なモダリティの情報を統合する役割を

担っているという点でその重要性が高い。

Transformer は、入力系列を受け取り符号化する encoder と、出力系列を生成する decoder の二つを組み合わせた Encoder-decoder モデルと呼ばれるモデル構成として提案されており、主に機械翻訳用のモデルとして利用され大きく発展してきた。その後、言語モデルに転用する際に、入力を decoder の前方文脈、いわゆるプロンプトとして扱うことで decoder のみの構成で用いる方法が主流となった。Transformer のモデル構成は、入力層に加えて、入力系列全体に対する位置情報を表す位置符号化または位置埋め込みを用い、系列全体を対象とした多頭注意機構（multi-head attention）と、位置ごとに独立な処理となる３層フィードフォーワードネットワーク、層正規化の３種の組み合わせを一連の処理とした多数の中間層で構成される。一般論として、多頭注意機構において入力系列内での情報の混ぜ合わせが行われ、３層フィードフォーワードネットワークの中に知識が埋め込まれていると解釈されることが多い。

Transformer はモデル構成として極めて簡潔な作りをしているが、パラメータ数や学習データ量を多くすることで強力なモデルとなる。パラメータ数や学習データを変数として性能をプロットした際に、概ね対数スケールに対して線形に性能が向上することが知られており、このような関係をスケーリング則（Scaling Laws）と呼ぶ。実際、OpenAI の GPT シリーズは、LLM のパラメータ数が、GPT[10]（1.17 億、2018 年）、GPT-2[12]（15 億、2019 年）、GPT-3[13]（1,750 億、2020 年）、GPT-4[14]（2023 年時点では、未公開であるが２兆と推測されている）のように大規模化の一途をたどってきた。スケーリング則は、性能向上のために膨大な学習データが必要であることも意味するが、LLM の場合、テキストの次の単語を予測させる問題設定とすることで、既存の膨大なテキストをそのままの形で学習データとして活用することができる。この段階の大規模な学習は事前学習と呼ばれる。また、LLM で用いられるように、人手による正解を付与する必要がないデータを活用して学習する方法を「教師あり学習」の中でも特別に「自己教師あり学習」と呼ぶこともある。

パラメータ数および事前学習コーパスが非常に大きくなり、モデルの学習が高コストになったため、効率的な学習方法が模索されている。また、小さなモデルで大きなモデルに近い性能を達成することを目的とした研究が行われている。Transformer 以外にも、Retentive Network[15]や、Mamba[16]のような状態空間モデル（State Space Model）などの他のアーキテクチャも提案されており、設定によっては Transformer と同等あるいはそれ以上の性能が報告されている。また、複数のモデルを組み合わせる MoE（Mixture of Experts）やモデルマージなどの手法についても盛んに研究が行われている。

最後に、生成 AI の発展の中で非常に重要な意味を持つファインチューニング（fine tuning）について説明をする。ファインチューニングとは、大規模データによって事前学習された生成 AI のベースモデルに対し、正確性・有用性や安全性を満たしたコンテンツを生成するように、あるいは特定のタスクやドメインに適応させる目的で、モデルを調整する手法である（事後学習とも呼ばれる）。具体的には、入力（プロンプト）とそれに対する望ましい出力のデータ、あるいは入力と回答候補に対して回答の適切さをア

ノテーションしたデータを用意し、ベースモデルに対して追加学習を行う。前者は指示チューニング（instruction tuning）、後者は人間のフィードバックによる学習（learning from human feedback）と呼ばれる。これにより、生成AIモデルは、プロンプトに対してどのように回答すべきか、非倫理的な質問に対してどのように反応すべきか、などを学習する。最初期のChatGPTはGPT-3というLLMをベースにしているが、オリジナルのGPT-3は人間の指示に対してその要求に合う回答を出力することや、非倫理的な入力に対して回答を拒否することが不十分であった。オリジナルのGPT-3に対して、人間の価値観や倫理観に合った出力を行わせるための技術としてファインチューニングが施され、これが非常に有効であったことがChatGPTの爆発的な利用拡大の引き金となった。LLMをはじめとして現代の生成AIモデルでは、この技術がほぼ標準的に適用されている。

⑵　各種のモダリティの扱い

人間は、視覚や聴覚などの五感から得た情報を処理することで、外界の状況を理解してコミュニケーションを行っている。言語や画像、音声など、情報伝達の手段（種類）のことをモダリティと呼び、複数のモダリティの情報を処理することをマルチモーダル処理と呼ぶ。マルチモーダル処理では、言葉での指示の通りに画像を生成する画像生成や、図や表の内容を言葉で説明する説明文生成など、有益な応用がある。

①　テキストの扱い

まず、基盤モデルがテキスト（言語）をどのように扱うかを説明する。文字の並び（文字列）として表現されるテキストをTransformerなどのモデルに入力するときは、単語もしくはサブワード（単語よりも小さい文字列）を単位としたトークン列に区切ることが一般的である。この処理は、トークナイゼーション（トークン化）と呼ばれる。単語をトークンの単位とすると、Transformerの語彙から漏れてしまう単語（未知語）が出てくるが、文字をトークンの単位とすると、トークン列が長くなりすぎて扱いにくい。サブワードは、単語と文字の中間くらいの単位で、BPE（バイト対符号化）やユニグラム言語モデルなどを用い、学習コーパスから統計的に語彙を構築する。

サブワードを学習するアルゴリズムは、言語に依存しないため、多言語のコーパスから語彙を構築することで、多言語に対応したトークナイゼーションを実現できる。ところが、語彙構築アルゴリズムは、学習コーパスの文字列の出現の統計に基づくため、コーパスにあまり含まれない言語から語彙が選ばれにくいことに留意すべきである。また、海外由来のLLMでUTF-8を文字コードとして採用し（アルファベットに対して自然な取り扱いである）、１文字１バイトを想定している場合、語彙に収録されていない日本語の文字はUTF-8のバイト列（通常は１文字３バイトであるため、１文字が３トークンに分割される）で表現され、不自然な取り扱いになることに注意が必要である。

基盤モデルのトークンとして、BOS（系列の冒頭）やEOS（系列の末尾）などの特殊トークンが使われている。以前は、タスクや言語を表す特殊トークンも使われること

があった。例えば、翻訳先の言語を表す特殊トークンとして、TO_EN（英語に翻訳）やTO_DE（ドイツ語に翻訳）などをつけることで、特殊トークンに対応する埋め込み表現が翻訳先言語を指定し、多言語翻訳を一つのニューラル機械翻訳モデルで実現することがあった。しかし、基盤モデルに指示チューニングが導入されたことにより、特殊トークンを用いなくても「英語に翻訳してください」という指示でタスクや言語を認識できるようになった。以降で説明するマルチモーダル基盤モデルで画像や音声の情報をトークン埋め込みとして与える場合、言語以外のモダリティの特殊トークンを導入していることに相当する。

プログラムコードについても、現時点ではテキストのモダリティの範囲で扱うことが一般的である。例えば、プログラムの仕様からコードを生成するタスクは、与えられた指示に対してコードの文字列を次単語予測で生成することによって実現される。また、ソースコードのデータで事前学習したLLMが高い推論能力を発揮するという報告があることから[17]、LLMの事前学習にThe Stack（GitHubのレポジトリからコメントとコードなどのペアデータを収集・抽出したデータセット）などのソースコードのデータを混ぜる事例が多く見られる。LLMのコード生成能力の評価には、HumanEvalやMostly Basic Python Problems（MBPP）などのベンチマークデータが用いられる。

事前学習コーパスに含まれない新しい情報への対応、ドメイン知識や社内文書などの特定知識の利用、さらに、ハルシネーション（事実と異なる情報を出力すること）への対策の一つとして、RAG（Retrieval-Augmented Generation）と呼ばれる技術の研究開発も盛んに進められている[18]。RAGでは、質問に対して、まず、ある文書集合を対象とした情報検索を行い、その上位数件の文書、もしくはその中のより関連する箇所をプロンプトとしてLLMに与え、その文書の情報を踏まえてLLMが回答する。すでに、このようなRAGのサービスを提供するITベンダーも多数存在する。

② 画像・映像の扱い

画像生成AIでも、人間が撮影・描画・制作したものと区別がつかない生成が可能になっている。こうした画像を「良い画像」とすると、純粋に乱数で発生したような画像は「悪い画像」と言える。画像生成AIには、様々なAIモデルが利用されてきた。具体的には、変分オートエンコーダ（Variational Autoencoder：VAE）[19]、敵対的生成ネットワーク（Generative Adversarial Network：GAN）[9]、拡散モデル（Diffusion Model）[20]が知られている。これらはいずれも、まず大量に集めた「良い画像」によりモデルを学習しておく。このモデルにより「よい画像」が生成可能となる。生成された画像は、学習時に使った画像とは異なっているが、人間が見れば明らかに「良い画像」の部類に入るものが生成される。これにより、画像が生成できるAIとして話題になっている。

VAEでは、画像を潜在変数に変換するエンコーダと、潜在変数を画像に変換するデコーダから構成され、エンコーダとデコーダを直結した状態で、学習用の画像を入力して、できる限りそのまま出力されるように学習する。学習されたモデルは、新たな

「良い画像」を生成できるが、解像度が落ちるなど品質に問題があった。GAN は、潜在変数から画像を生成する生成ネットワークと、「良い画像」と「悪い画像」を識別する識別ネットワークから構成され、相互に敵対的に学習させることにより、VAE よりも高品質の「良い画像」が生成できるとして話題になったが、学習が安定しないという問題があった。

現在、画像生成 AI として最も期待されている技術が、拡散モデルである。この AI では、画像が熱拡散過程に従い徐々に正規分布に従うノイズに変化していく過程を考える。そしてそれを逆変換しノイズを除去して元の画像に徐々に近づけていくネットワークを学習することにより実現している。GAN と比べて学習は安定しており、生成される画像の品質も高い。最近では、ControlNet[21]など、よりきめ細かい要望に応えるための技術の検討が盛んに行われている。

VAE、GAN、拡散モデルのいずれも、「良い画像」の従う確率分布に従う画像を乱数的に生成する。画像の表現を考えると、例えば 16×16 画素、各画素二階調の極めて単純な画像ですら、「良い画像」も「悪い画像」も含めて 2 の 256 乗、すなわち 10 の 77 乗以上の異なる画像がある。拡散モデルの一つ Stable Diffusion は、数十億画像で学習したと言われているが、10 の 9 乗に過ぎない。そう考えると、生成される画像は学習画像の類似や模倣となりかねないのも、手法の特性上やむを得ないと考えられ、学習画像とはあらゆる観点から全く異なる「良い画像」の生成は、原理的に困難と考えられる。

このように画像生成 AI には Transformer とは異なるモデルが利用されてきたが、昨今では基盤モデルとしての画像エンコーダにも Transformer をベースとしたモデルが用いられることが珍しくなくなってきた。例えば、拡散モデルにおけるノイズ除去には、従来 U-Net が用いられてきた[20]のに対し、最近では Diffusion Transformer が良好な性能を示すことが確かめられたことで[22]、画像生成においても Transformer をベースとしたモデルが用いられるようになった。結果的に、あらゆる入出力を扱う基盤モデルや生成 AI の研究開発領域において、Transformer をベースとしたモデルが最重要技術の一つと捉えられるようになっている。なお、こうしたモデルでも、先述した学習画像とはあらゆる観点から全く異なる「良い画像」の生成が原理的に困難という点は免れ得ない。

最近は Sora[3]など、動画を生成する AI も登場し、その生成された映像の品質に皆驚いた。これらも基本的には拡散モデルに基づいた手法であり、二次元の画像表現に対し、時間軸も含めた三次元の動画表現を考えた拡散モデルにより、動画を生成している。今のところ生成される動画は短いもののみだが、今後はよりストーリー性を持つ動画の生成などへの展開が期待される。

[3] https://openai.com/index/sora/

③　言語と画像のマルチモーダル処理

言語、画像、音など、異なる形式のデータを組み合わせて利用することも可能になってきた。上述の通り、これをマルチモーダル処理と呼ぶ。我々人間は、マルチモーダル処理を行っている。例えば、ウェブページを閲覧する際には、そこに書かれている言語に加え、添えられている画像を見て、それらを総合的に判断して内容を理解している。以前のマルチモーダル処理では、異なるモダリティからの情報を単純に組み合わせ、よりリッチな情報として利用する程度であった。これに対して、最近のAI技術の進歩により、異なるモダリティを「結びつける」技術が多数提案されている。例えば、条件付拡散モデル（Conditional Diffusion Model）は、画像の内容を説明する文章（プロンプト）を与えることで、人間の要望に応じた画像を生成する技術である。例えば「A photo of dog」というプロンプトを与えれば、犬の写真のような画像が得られる。これは、あるモダリティ（言語）を別のモダリティ（画像）に変換する技術とも言える。プロンプトについては「A high resolution、 masterpiece quality photo of dog」などと与えた方が、良い品質の画像が得られるなど、場当たり的な知見も散見される。

関連して、異なるモダリティであっても、それが本質的に同じものであれば、同一のデータとして表現する技術も開発されている。例えば、CLIP[23]と呼ばれるモデルでは、犬という単語と犬の画像を、ほぼ同じ特徴ベクトルとして表現する。これにより、コンピュータにとっては、同じ事物を指す単語と画像を区別する必要がなくなる。なおCLIPは、拡散モデルにおいてプロンプトを条件として画像生成する際に、プロンプトのエンコーダとしても用いられる。

言語と画像のマルチモーダル処理のための基盤モデルとしては、様々なアーキテクチャが提案されている。テキストから特徴量を取り出すテキスト・エンコーダ、画像の特徴量を取り出す画像エンコーダ、画像とテキストの特徴量を結びつける変換器、画像とテキストの情報を統合する機構で構成されることが多い。こうしたアーキテクチャの代表例であるLLaVA[24]は、言語エンコーダにVicuna（Llamaの派生モデル）、画像エンコーダにCLIP、画像からテキストへの特徴量変換器に線形層を用い、画像の特徴量ベクトルをテキストのトークン埋め込みと同様に扱い、Transformer（Vicuna）の自己注意機構で画像と言語の情報を統合している。LLaVAは、まずテキストと画像のエンコーダのパラメータを固定し、画像とテキストが対になった学習データで特徴量変換器の線形層のみを学習し、続いて画像とテキストを含む対話データや質問応答データを用い、ファインチューニング（指示チューニング）を行うことで構築される。

④　音声・音楽の扱い

音の生成AIは、ここ数年で大きく進化した。従来からテキストを読み上げるテキスト音声合成システムなどは存在していた。最近は、この従来通りの生成タスクに加えて、ユーザーが所望する内容を記述したプロンプトから効果音、環境音、音楽などを生成するといった多種多様な音の生成モデルが提案・検討され、一部は既に実用化

されている。

人間の音声、効果音、環境音、音楽ではその音の特徴は大きく異なる。しかし、これらの生成技術を発展・加速させているのは、「音のトークナイゼーション」という共通の基盤技術である。音のデジタル信号は、標本化周波数に応じた点の時系列で表現され、各点は連続値を持つ。この長い連続時系列で表現された音の波形を短い離散記号列に自動変換し、また元の波形を再合成する技術が音のトークナイゼーション技術であり、言葉の音素的特徴を重視した HuBERT 手法[25]から、音の細部まで再合成することを重視した Encodec 手法[26]や SoundStream 手法[27]まで色々な方法が提案されている。

これらのトークナイゼーション手法を用いれば、人間の音声、効果音、環境音、音楽といった音の特徴に関係なく、音データは離散記号列に変換される。その結果、音のトークンをテキストのトークンと同様に扱うことが可能になり、言語モデルと全く同様の事前学習が可能になる。また、ユーザーが所望する音の内容を記述したプロンプトを条件変数とした音の言語モデルを通常の言語モデルと同様に学習することも可能になる。

このような音のトークン化と言語モデルの組み合わせは強力であり、Google 社および Meta 社は、プロンプトから多種多様な音を自動生成するモデルをそれぞれ AudioLM[28]、AudioGen[29]という名前で発表し、またプロンプトから音楽を生成するモデルをそれぞれ MusicLM[30]、MusicGen[31]という名前で発表した。これらを発展させた商用サービスも既に実用化している。Stability AI 社は、ユーザーがアップロードした音や音楽を、ユーザーが記述したプロンプトを基にして所望の音へ変換する Stable Audio というサービスを提供開始した。Suno 社と Udio 社は、ユーザーが入力した歌詞から音楽を生成するサービスを開始した。日本語を含む複数の言語による歌詞入力に対応している。その一方、この２社はアーティストの楽曲を無断使用したという著作権侵害を理由に裁判を起こされている[4]。

入力テキストを読み上げる従来のテキスト音声合成も音のトークン技術で大きく進化した。具体的には、従来は扱いにくかった笑い声などの非言語音声を扱うことも可能になり、対話調の音声もよりリアルに自然発声風に合成することが可能になりつつある。

⑤ 生成 AI のロボティクス応用

LLM によるロボットへの対話指示の理解に関しては、Google の Say-Can[32]など数多くの研究事例がある。言語理解から動作計画へ至るプロセスでは大きな性能向上が実現され、特に移動ロボットのナビゲーション指示に利用されている。しかし、物体操作などのより複雑な動作計画を具体的な運動に変換することは容易ではない。

言語側からではなく、生成 AI を運動スキルに利用する研究も盛んになっている。

[4] https://www.bbc.com/news/articles/ckrrr8yelzvo

例えば、RNN（Recurrent Neural Network）などのダイナミクス生成モデルを歩行ロボットの強化学習に用いるアプローチ（例えば[33]）は、極めて大きな進展を遂げている。これらの研究ではシミュレーションで学習するが、ロボット本体のみを精度良くモデル化し、実環境とは大きく異なったより過酷な仮想環境で膨大な試行回数を学習させる方法が効果を挙げている。

しかし物体ハンドリングに関しては、対象物体である柔軟物や粘性流体などのモデル化が必要であり、シミュレーション学習のアプローチは容易ではない。そこで実世界で人間がロボットを操作することで動作データを集め（模倣学習）、LLM をファインチューニングする試みが進んだ。Google の RT-1、RT-2、RT-X が代表研究である。特に RT-X[34]は世界 33 研究機関から 100 万エピソードデータを収集し、言語と運動のスムーズな変換がある程度実現された。しかし集めたデータは、単一マニピュレータによる Pick and Place タスクが中心であり、複雑な作業は実現できていない。

マニピュレーションの学習モデルとして、RNN ベースの深層予測学習（例えば[35]）、Transformer ベースの ACT（Action Chunk Transformer）[36]、Diffusion Model ベースの Diffusion Policy[37]などが提案されている。これらの End-to-End の生成 AI モデルは、多点接触を伴う双腕協調や多指ハンド操作などより熟練した人間の動作を、一定量の学習データで模倣学習させることが可能である。Google DeepMind の ALOHA[36]、Tesla の Optimus、OpenAI&1X の EVE など、多くの双腕ロボット（人型ロボット）で、“ロボット基盤モデル”のための大量の動作データ収集の競争が展開されている。この過程で改めて力覚と触覚のモダリティの重要性が再認識されている。

(3) 生成 AI に対する規制の動き

このように急速に発展する生成 AI に対して、規制枠組の整備も各国において活発に進められている。その手法や規律強度は様々ながら、主として１．影響力の大きい基盤モデルの開発者に焦点を当てた規律、２．ディープフェイクや偽・誤情報がもたらすリスクへの着目、３．それらが流通するソーシャルメディアなどでの対応という３点に特徴づけられる。

2021 年４月に欧州委員会から出された欧州連合 AI 法の当初提案は、AI システム全般について、その用途により、許容できないリスク、ハイリスク、限定的リスク、最小リスクの４段階に分類した規制枠組を構築するものであった。そこで念頭に置かれているのは、AI が組み込まれた製品が、人の生命・身体や財産に危害を及ぼすリスクと、AI による個人データの自動処理に基づく決定がもたらす個人の権利・利益へのリスクであった。それが 2022 年末の ChatGPT の登場を受け、審議過程の中で生成 AI を含む汎用目的 AI モデル（General Purpose AI Model）についての規律枠組が加えられた。具体的には、汎用目的 AI モデルの提供者（開発者など）全般に技術文書作成や透明性義務を課すほか、訓練に使用される計算量が 10^{25} FLOPs を超えるなどの閾値を満たす汎用目的 AI モデルについて、偽・誤情報の大規模な伝播を含む広範なシステミックリスクを特定し、それらを軽減することなどを義務づけた。このほか、汎用目的 AI モデルを含むコンテンツ

生成AIシステムの提供者は、アウトプットが機械可読な形でマーク付けされ、人為的に生成または操作されたことを検知可能にすることなどが義務づけられた。これはソーシャルメディアなどの運営事業者に対して偽・誤情報などへの対応を義務づける、デジタルサービス法（Digital Services Act）の規律を補完する位置づけにある。

米国は、バイデン政権下の2023年10月に「人工知能の安全かつ信頼性の高い開発および利用に関する大統領令（Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence）」を発令し、CBRN（Chemical、Biological、Radiological、and Nuclear、化学・生物・放射能・核）兵器開発を容易にするなど、安全保障などに影響を及ぼし得る大規模なデュアルユース基盤モデル開発者に対して、安全性テストの結果や開発・製造に関わる情報の政府への報告を義務づけたが、トランプ大統領就任直後の2025年1月に同大統領令は撤回された[5]。他方で、州法レベルでの立法作業は進展しており、特に多くのAI企業が本拠地を置くカリフォルニア州では、2024年９月、AI生成コンテンツのラベルづけや、アーティストの肖像や声の保護、性的に露骨なディープフェイクの規制、選挙におけるディープフェイクへの規律やソーシャルメディア側での対応、生成 AI の訓練に用いられたデータ概要の開示などを定める多数のAI関連法案が成立している[6]。

英国においては、2023年３月の政府AI政策白書「AI規制：イノベーション志向のアプローチ（A pro-innovation approach to AI regulation）」の中で、EUのようなAIに特化した広範な立法を行うことはせず、既存法の枠組みの中でのリスク対応を進める方針を示していた[7]。しかし、労働党新政権樹立後の2024年７月に行われた国王施政方針演説の中では、「最も強力なAIモデル」の開発者に対して適切な法的要件を確立するための立法を目指すことが明示された[8]。今後立法作業が進められる見込みである。

日本はこれまで、AI規制に関しては、AIに特化した立法を行う選択肢はとらず、総務省と経済産業省が取りまとめた AI 事業者ガイドライン[38]に代表される非拘束的なソフトローでの対応を重視してきた。しかし、2024年８月には内閣府AI戦略会議のもとにAI制度研究会が設置され、EUや米国の制度的対応を参照しつつ、日本におけるAI制度のあり方についての検討が進められ、2025年２月には、政府の司令塔機能の強化と安全性確保などを中心とした速やかな法整備を求める中間とりまとめが公表された[9]。また、2024年９月には、総務省の「デジタル空間における情報流通の健全性確保のあり方に関する検討会」が、生成AIが生み出す偽・誤情報の流通へのSNSなどのソーシャルメディア側での対応を視野に入れたとりまとめを公表する[10]など、具体的な制度のあり方についての議論が進められている。

---

[5] https://www.whitehouse.gov/presidential-actions/2025/01/initial-rescissions-of-harmful-executive-orders-and-actions/

[6] https://www.jetro.go.jp/biznews/2024/09/a5ec5ef37532a9d9.html

[7] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1176103/a-pro-innovation-approach-to-ai-regulation-amended-web-ready.pdf

[8] https://forbesjapan.com/articles/detail/72475

[9] https://www8.cao.go.jp/cstp/ai/index.html

[10] https://www.soumu.go.jp/main_content/000966997.pdf

３　生成 AI の脅威と課題

生成 AI は、人々の知的作業の効率を高め、人々の生活を便利にし、産業を成長させるというプラスの面が大いに期待されている。一方で、様々なリスクをもたらし得るという懸念が指摘されている[39][40]。

本章では、生成 AI を受容・活用する社会の実現に向けて、どのようなリスクに対処していく必要があるかをまとめる。まず、生成 AI がもたらす脅威として、現状で一般的に把握されているものを、生成 AI によって生じる現象に着目して列挙する。次に、社会における法的・倫理的なリスク、特に著作権侵害、名誉毀損、肖像権侵害などについて整理する。また、生成 AI を開発するためにもリソースなどの障壁があり、その点についてもまとめる。最後に、生成 AI を受容・活用する社会の実現に向け、生成 AI モデルが備えるべき要件を整理する。

(1)　生成 AI の脅威

生成 AI が世の中に公開され利用され始めると、役に立つ要素だけではなく、人間の活動に対しての脅威になり得ることが分かってきた。例えば、一部の専門家だけしか知り得るべきではない危険な情報へ比較的容易にアクセスできるようになることや、虚偽の情報が流布され社会が大きく混乱するなどの脅威である。それらをここでは社会的なリスクとして整理する。ただし、生成 AI は文字通り日進月歩で進んでいることに注意を払う必要がある。現在脅威と認識されていることがある技術で解決されたり、現在は全く見えていない脅威が、ある日突然認識されることは十分にあり得る。

①　ハルシネーションによる脅威

生成 AI として最初に公開された言語モデルは、様々な言語による質問に対して、何らかの回答を流暢に言語で返すことで、世の中に衝撃を与えた。流暢性や関連性、情報量については一般的に高い評価を受けている。しかし、情報の正確性については、まだ完全ではないことが分かっている。公開されている百科事典に記載されていることについても正確な回答を生成できないことなど、様々な間違いが認められる。このような間違いはハルシネーションと呼ばれている。ハルシネーションが起こる原因として、言語モデルが仕組みとして次の単語を確率的に生成することや、事前学習コーパスに含まれ得る誤情報などが挙げられる。商用のシステムにおいても、回答が不完全なので注意して利用することという利用規約があることが一般的である。現状の技術では解決が困難な問題であるが、その対策のための研究は、世界的に活発に取り組まれている。

②　高品質な生成メディアによる脅威

生成 AI の進展により、顔、音声、自然言語などの人間由来の情報を大量に学習することで、本物と見紛う画像・映像、音声、文章といった高品質な生成メディアの自動生成が可能になりつつある。生成メディアは、マーケティングや広告、教育、コンテ

ンツ制作など多岐にわたる分野で活用されており、生成メディアの品質自体も向上している。一方で、生成メディアの負の側面として、詐欺や思考誘導、世論操作を行う目的で、愉快犯や攻撃者がフェイク画像・映像、フェイク音声、フェイク文書といったフェイクメディアを生成・拡散させており、社会問題となっている。例えば、AI で生成したフェイク音声によって英国企業の幹部に「なりすます」ことで現金を搾取した事例[11]や、ウクライナ大統領のディープフェイクによるウクライナ国民への降伏呼びかけ[12]などが生じており、国内においても、静岡県の水害被害に関するフェイク画像の拡散[13]や、岸田前首相のディープフェイクの拡散[14]など、フェイクメディアによる詐欺や思考誘導、世論操作は現実の脅威となっている。対策として、メディアの詳細解析によって不自然な部分や不整合を見つけ、フェイクや改ざんを検出する技術や、電子透かし、証明書、ブロックチェーン（ハッキングや詐欺が難しい方法で、取引の記録を複数の場所に分散して管理する技術）などを用いて、出所や経路を検証可能にする技術が活発に研究されている。

③　非社会的な回答やアクション生成による脅威

正確性の問題だけではなく、生成 AI が一部の専門家しか知るべきではない危険な情報に属する発言をする危険性も知られている。そのような危険発言をしないように、指示チューニングによってガードがかけられているものの、そのガードを破って危険発言をさせるプロンプトインジェクションや脱獄（ジェイルブレイク）と呼ばれる行為が問題視されている。特に欧米では CBRN に対するセキュリティに重点が置かれ、CBRN に関する情報の保護に大きな注目が集まっている。また、社会的なバイアスが認められる発言、差別的な発言については、組織的な社会扇動にもつながる可能性がある。例えば選挙の操作、一部の人に対する社会的利益・不利益を生じさせる可能性が指摘されている。アダルト情報、ヘイトスピーチなど反公序良俗についての発言も社会的な脅威となり得る。また、違法行為への加担や非倫理的行為への加担も、生成 AI によって可能となることも知られている。これらの社会的安全性に関する問題に適切に回答することは、現在の生成 AI には難しい課題である。

また、生成 AI は目標が与えられると、それを達成する手順（副目標の系列）に分解してアクションを実行することが可能である。このとき、不適切な副目標が生成されてしまう可能性がある。例えば目標達成に邪魔なものを排除したり、停止スイッチを無効にしてしまったりといった不適切なアクションが生成され得る（道具的副目標収束の問題といわれる）。そのようなケースを回避するようにガードをかける必要があるが、網羅的にガードをかけることは難しい。

---

[11] https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video
[12] https://www.nikkei.com/article/DGXZQOGN177EW0X10C22A3000000/
[13] https://www.yomiuri.co.jp/national/20220927-OYT1T50208/
[14] https://www.sankei.com/article/20231114-LLOVR22LSNOVNFWVGOIRN5JIBU/

④　機密情報漏洩による脅威

生成AIが国家、企業、個人の機密情報漏洩をしてしまう危険性も知られている。情報漏洩が生じる原因にはいくつかの可能性がある。まず、モデル構築のために収集された学習データの中に機密情報が含まれる場合が考えられる。これは、そのような情報をインターネットなど何らかの方法で取得可能な状況に置かないという国家や企業の規制が求められる。

また、多くの生成 AI サービスは入力された質問を学習のために再利用しているため、ユーザーが機密情報を質問の一部として入力した場合に、そこから機密情報が漏洩する危険がある。生成 AI サービス利用時に入力を再利用させないよう設定することや、生成AIサービス事業者との契約でそれを禁止することもできるが、それだけでは機密情報の保護として十分とはいえない場合もある。生成 AI モデルで機密情報を扱わざるを得ない場合は、組織内でモデルの運用を行い、さらにその利用範囲を限定することが考えられる。

⑤　ハッキングによる脅威（サイバーセキュリティ）

コード生成 AI によってハッキングプログラムが作成され、そのハッキングプログラムを使って、道路管制システムや銀行システムのようないわゆる重要インフラに致命的な打撃を与える可能性がある。もちろん、コード生成AIを用いて、上記のような意図でハッキングプログラムを使用した場合には、電子計算機損壊等業務妨害罪などの犯罪に問い得るが、刑罰による威嚇だけでは重要インフラへの打撃を止めにくい点が、社会的リスクとして存在する。

また、AI特有の脆弱性を突く攻撃として、学習データや入力データに細工して、誤認識や想定外動作を誘発する攻撃（ポイズニング攻撃、バックドア攻撃、敵対的サンプル攻撃など）や、モデルやデータを搾取する攻撃（モデル抽出攻撃、モデルインバージョン攻撃、メンバーシップ推測攻撃など）が知られており、これらは生成AIが組み込まれたシステムにおいても脅威となる。

⑵　生成AIの法的・倫理的懸念

生成 AI モデルの運用に伴うリスクについて、法的リスクおよび倫理的懸念を取り上げる。両者は重なり合う場合もある。

①　著作権侵害

著作権侵害については、生成AIの開発・学習段階と生成・利用段階の各段階において、異なるリスクが存在する。

生成・利用段階においては、既存の著作物に依拠する形で（依拠性）、その著作物に類似したもの（類似性）を生成・利用する場合に著作権侵害と判断される可能性がある。学習用データに既存の著作物が含まれているだけで依拠性があると評価してよいかについては議論がある。

開発・学習段階においては基本的には著作権侵害にはならないと考えられている。しかし、著作権侵害となるような利用を目的として開発・学習を行う場合には著作権侵害とされる可能性があり得るほか、著作物たるある作品のデータを用いて機械学習をさせること自体が著作権侵害であるという見解もある。アメリカでは既に訴訟が行われている[15]。

② 名誉毀損・信用毀損

ハルシネーションにより個人の情報が歪曲されて広まった場合に名誉毀損のリスクが生じる。実際にオーストラリア南部ヘップバーンシャーの市長が、ChatGPT に自身が贈賄罪で服役していたという虚偽の説明があるとして、訂正されなければ開発した OpenAI を名誉毀損で訴える、と述べたと報道されている[16]。

日本法においても、個人または法人に関する虚偽の事実の摘示は、その事実が当該人の社会的評価を低下させるものである場合、名誉毀損に基づく不法行為や、果ては名誉毀損罪ないし信用毀損罪にも問われる可能性がある。

確かに、ChatGPT が、情報を自ら拡散させるわけではないが、ChatGPT に特定の質問をすると、虚偽の事実を含む回答が返ってくることについて再現可能な場合には、データベースに対してアクセス可能になっていることにより、公然の名誉侵害であると評価される可能性が捨てきれない。

上記オーストラリアの市長の記事においては、ChatGPT 上に誤情報があった、と指摘されており、事実に反するデータの速やかな訂正や、必要に応じて、情報が訂正された旨の注記を入れるなどが可能になるのであれば、大きな紛争には至らないものとも思われる。ただし、訂正に必要な立証活動のあり方などについては、法的な問題としてあらかじめ解決しておく必要がある。

また、2023 年には、アメリカの連邦取引委員会が、ChatGPT が風評被害を引き起こす可能性のある虚偽情報を提供したという疑惑について追及を行っている[17]。

③ 肖像権・パブリシティ権侵害

パブリシティ権の侵害というためには、専らその保護客体の有する顧客吸引力の利用を目的とすると言えることを要する。パブリシティ権の保護客体になるものは本人の人物識別情報として広く解される可能性がある。例えば、「声」もパブリシティ権の保護客体に含まれる、という理解もあり（「AI 時代の知的財産権検討会」中間とりまとめ[18]56 頁）、これによれば、ある歌手の声を生成 AI で生成して別な歌唱をさせるような「AI カバー」における「声」の部分にパブリシティ権の保護が及び、「声」の利用態様（「専らその保護客体の有する顧客吸引力の利用を目的とするといえる場合」

[15] https://www.bbc.com/japanese/67831445
[16] https://jp.reuters.com/article/life/-idUSKBN2W308P/
[17] https://www.forbes.com/sites/tylerroush/2023/07/13/ftc-investigating-chatgpt-maker-openai-for-providing-false-information-report-says
[18] https://www.kantei.go.jp/jp/singi/titeki2/chitekizaisan2024/0528_ai.pdf

か否か）によっては同権利の侵害になる可能性がある。アイドルを生成AIでモデル化して会話できるようなソフトもそのアイドルの肖像権（自らの肖像をみだりに利用されない権利）やパブリシティ権の侵害のリスクが生じる[19]。実在の人物の容貌と生成された人物肖像が類似する場合には、肖像権・パブリシティ権の対象となる可能性があるが、議論はまだ定まっていない。

④ コード生成AIに関する法的リスク

権利侵害や犯罪といった違法行為の道具として用いられるプログラムは、多くの場合、適法な用途も考えられないわけではなく、そうしたプログラムはデュアルユースツールと呼ばれることがある。仕様を与えるとプログラムコードを作る生成AIであるコード生成AIも、そのようなツールの一種だと考えられる。違法行為に利用する目的をもってコード生成AIを用いてプログラムコードを作り、これを違法な用途に利用した場合には、現実の利用の時点で、利用者は犯罪や権利侵害をしたと評価され得る。例えば、コード生成AIによりSNSなどのいろいろなメディア情報を組み合わせて、個人を特定するようなプログラムが作られる可能性があり、そのようなプログラムにより、例えば未成年犯罪者が容易に特定されてしまうなどした場合には、プライバシー侵害と評価されるリスクがある。

これだけをみると、コード生成AIが、違法行為を助長しているようにみえるかもしれない。とはいえ、違法行為に使われる道具の開発を助長したという「幇助」の構成で、共同不法行為や幇助犯という観点からコード生成AIの開発者に対して法的責任を負わせようとするのは行き過ぎかもしれない。コード生成AIによって生み出される無数の社会的に有益なプログラムの可能性も踏まえて考えるべきであり、規制の当否について法的な議論を尽くすべきである。

⑤ コンテンツ利用規約違反

インターネット上のコンテンツを利用する場合、「生成AI学習への利用を禁止する。」というような規約が書かれていたにも関わらず、その規約に違反した利用を行えば、訴訟につながるおそれがある。

例えば、読売新聞の読売新聞オンライン利用規約[20]の第７条第６～８項においては、以下の禁止事項を規定しており、そのことを知らずにクロール（ロボット検索）などでコンテンツを学習に利用し、規約に違反した場合に訴訟につながるおそれがある。

- データマイニング、テキストマイニング等のコンピュータによる言語解析行為
- 当社コンテンツを、クローリング、スクレイピング等の自動化された手段を用いてデータ収集、抽出、加工、解析または蓄積等をする行為
- 生成AI等（人工知能、検索拡張生成、RPA、ロボット、プログラム、ソフトウェア

[19] https://ecaiz.com/企業が生成AIで作ったもの使うときに注意すべき７/
[20] https://www.yomiuri.co.jp/policy/terms/

を含むが、これらに限られない）に学習させる行為（検索等の利用により検索エンジンの生成 AI 等が結果的に学習することとなる行為を含むが、これらに限られない）または生成 AI 等を開発する行為

⑥ 芸術的活動への脅威

生成 AI によるアーティストのモデル化は、アーティストが作ってきた作品とは具体的な表現においては共通せず、アイデアのレベルにおいてしか共有しないものとして、著作権法における「類似性」は認められない可能性がある。しかし、「類似」と認められない場合でも、そのアーティストのテイストが感じられる作品を作りだす可能性がある。アーティストの創作意欲を大いに削ぐ可能性があり、また、法的リスクとしてパブリシティ権侵害の可能性もある。

⑦ 社会の価値観や文化の変化

死者と生成 AI で会話できるようなソフト[21]は、人間の死生観に大きな影響を与える。また、法的リスクとして、死者の名誉毀損になる可能性もある。

生成 AI の出力は、学習データの傾向や指示チューニングの過程に内在する価値観・倫理観やバイアスを反映したものになる。したがって、ある生成 AI を使い続けたり、ある生成 AI の回答に依存しすぎたりすると、人々や社会の価値観や文化にも影響を与えると考えられる。

(3) 生成 AI モデル開発の障壁

生成 AI モデルの開発には、膨大な計算資源とデータを要する。2012 年に深層学習が注目されるきっかけとなった AlexNet による ImageNet Large Scale Visual Recognition Challenge（ILSVRC）では、１台あたり約７万円の GPU２台を用いて５日かけて学習が行われた。この 10 年後の 2022 年に OpenAI が行った GPT-4 の学習の詳細は公開されていないが、１台あたり約 100 万円の GPU を 25,000 台用いて 90 日かけて学習したと予想されている[41]。OpenAI GPT-4 の１回の学習に要した計算費用は 7,800 万ドル、Google Gemini Ultra では 1 億 9,100 万ドルに及ぶとの推定もある[42]。深層学習モデルの開発に用いられる計算資源量は、2012 年までは概ね半導体の進歩と同じ速度で２年に２倍のペースで増加していたが、2012 年を境に、このトレンドは半年で２倍に変化している[43]。同じ価格で利用できる計算資源が、２年で２倍（Moore の法則）にしかならないため、半年で２倍、つまり２年で 16 倍のペースを実現しているということは、２年毎に投資額が８倍になっていると概算できる。2019 年６月に Microsoft は OpenAI に 10 億ドルの投資を行い[44]、2023 年１月にはさらに 100 億ドルの投資を行っている[45]ことからも、この傾向は顕著である。

このような莫大な投資が行われている背景には、生成 AI プラットフォームの覇権を

[21] https://www.itmedia.co.jp/business/articles/2404/27/news029.html

握るための競争と、投資に対する効果がべき乗則（スケーリング則）にしたがって定量的に予測できる[46]という二つの原理が働いている。このスケーリング則は、現行のTransformerモデルとCommon Crawlベースの学習データを改善することなく、モデルとデータを単純に増大させることで、性能がべき乗則にしたがって向上することを経験的に示したもので、得られる性能の下限を保証していると考えることができる。今後、モデルやデータが改善されれば、当然スケーリング則の予測よりも良い性能が得られるが、下限が保証されていることは投資を考える上では圧倒的な好材料となる。

しかし、スケーリング則があるからといって、学習データと計算資源をスケールアップさせていくだけで生成 AI モデルの開発がうまくいくわけではない。何千億ものパラメータを持つ生成AIモデルを何千、何万ものGPUを同時に使って、短期間で学習を終わらせるためには、分散並列学習手法やそれを実装したフレームワークとそれを実行するための環境構築に関する専門知識が必要になる。1年に1回しか故障しないGPUでも365台で学習していると、1日に1回どれかが壊れて学習が停止することになる。学習に用いるソフトウェアは非常に複雑でバグが混入していることも多々あり、GPU の数が一桁増えると顕在化する新たな問題もある。それらの対処法は論文などに書かれていない場合が多く、独自のノウハウが必要となる。さらに、1回に何億円もかかるような大規模な学習は何度も行うことができないため、試行錯誤をすることすら許されない。そこで、最先端の生成 AI モデルを開発した際に米国の大手企業が出した論文を参考にすることで、独自に行う試行錯誤の回数を減らすということが考えられる。しかし、論文に書かれている数字が間違っている場合があり、それを鵜呑みにして大規模な学習を行って全く性能が出ないモデルができあがったという事例が実際に起こっている[47]。しかも、大規模なモデルで初めて起きる問題であったため、小規模なモデルでの検証で問題を発見することはできなかった。

大規模な生成 AI の開発において生じる問題は、AI（人工知能分野）に関する知識とHPC（高性能計算）に関する知識の両方がないと解決できない場合が多い。しかし、このAIとHPCの境界領域の人材が圧倒的に不足しており、両分野に習熟した人材の育成は喫緊の課題である。

２章および本節で述べたように、モデルの学習には大規模なデータが必要となるが、当然のことながらデータには一定の品質が求められる。大規模なデータは、Common Crawlのようなオープンな Web アーカイブから構築されることが多い。例えば、Common Crawl全量から品質の良い英語テキストを抽出することによって構築されたコーパスであるFineWebデータセットは、15兆トークンからなる。Meta社のLlama3モデルも同規模のコーパスで学習されている。しかし、Web といえどもコーパス量には限りがあり、これ以上に大規模な学習データを構築し、データの規模によってモデルの性能を上げることが限界を迎えつつある。

この障壁を乗り越えるためには、オープンサイエンスや学術論文などのオープンアクセスの動きとの連携が重要である。また、オープンではない著作物（書籍、映像など）を利用することも有望であるが、著作者の権利に配慮しつつ、慎重に利用方法を検討し

ていく必要がある。さらに、医療・教育・行政サービスなどのデータを扱うことを視野に入れた、プライバシーやセキュリティに配慮したデータインフラの整備も重要である。一方、コーパスを人工的に自動生成して学習に用いることも研究されているが、まだその有効性は定かではない。

本節で述べた障壁を克服し、持続的で健全な生成 AI モデルの研究開発が行われるためには、オープンな活動と国際連携が重要である。OpenAI、Google などのモデルはその詳細や学習データが明かされないブラックボックスモデルであるが、オープンな生成 AI モデルの研究開発も進みつつある。我が国においては国立情報学研究所（NII）が主宰する LLM-jp[22]の活動があり、世界的には、米国 Allen Institute for AI （Ai2）の OLMo[23]、UAE の Mohamed bin Zayed University of Artificial Intelligence（MBZUI）を中心とする LLM360[24]などの活動がある。また、基盤モデルに関する国際連携の活動として Trillion Parameter Consortium（TPC）[25]がある。当初は、科学技術のための 1T パラメータ級の基盤モデルを構築することを目標としていたが、最近では、科学に関する難易度の高いベンチマークデータを作ることへの方針転換が検討されている。

(4) 生成 AI モデルの備えるべき要件

生成 AI モデルは、人間による指示（プロンプト）に基づき画像や自然言語テキストなどのコンテンツを生成する。AI システムの一種であるため、従来から AI システムに対して求められていた要件はそのまま当てはまる。ただし、従来の AI システムとは異なり、生成 AI モデルについては、出力（生成されるコンテンツ）が無制限で、あらかじめ限定することが不可能である。さらに、生成 AI モデルは汎用的なシステムであるため、利用される環境や文脈を限定することができない。よって、生成 AI モデルが満たすべき要件の定義はより困難になる。

また、生成 AI モデルはそのまま利用されるだけでなく、特定のドメインやアプリケーションに組み込まれて利用されることがある。例えば LLM に基づくチャットシステム（ChatGPT など）は、直接ユーザーとやりとりする利用方法に留まらず、医療ドメインや法ドメインなど特定ドメインでの利用や、ロボット制御システムやアンケート分析システムなど特定サービス・アプリケーションに組み込まれた利用も想定される。生成 AI モデルが備えるべき要件は、ドメインやアプリケーションに依存して変わるため、要件を一律に定義することはできない。生成 AI モデルの開発者、生成 AI モデルを利用したサービス・アプリケーションの開発者など、それぞれの立場や文脈で具体的な要件を定義する必要があることに注意すべきである。

以下では、生成 AI モデルが備えるべき要件について一般論を説明する。生成 AI モデルは人間の知的活動を模倣し、人間の知的活動をサポート・代替していくものであるた

[22] https://llmc.nii.ac.jp/
[23] https://allenai.org/olmo
[24] https://www.llm360.ai/index.html
[25] https://tpc.dev/

め、人間との関わりなしに議論することはできない（人間中心の AI）。したがって、生成 AI モデルは人間や社会が安心・信頼して利用できる必要があり、そのためには複数の要件がある。

① 正確性・有用性

生成 AI モデルにおいてまず重要な要件として、正確性・有用性が挙げられる。生成 AI は人間の指示にしたがってコンテンツを生成することから、正しい情報を指示に正確に従って生成すること（正確性）、および人間の要求に沿った有用な情報を生成すること（有用性）が求められる。ただし、生成 AI モデルの出力について、一般に唯一の正解は存在せず、正確性・有用性の定義は、ドメインや文脈、アプリケーションによって変わり得る。例えば、「日本で一番高い山を教えて」という指示に対し、「富士山」と答えるのは正確であるが、それが正確性・有用性を満たすと言えるかどうかは議論の余地がある。この生成 AI モデルが ChatGPT のようなチャットシステムである場合には、答えのみを出力することは有用性が低く、より多くの情報を提示することが望ましい。例えば「富士山は山梨県と静岡県にまたがる場所にある独立峰で、高さは 3,776m で、日本で一番高い山です。」といった根拠や関連情報を提示することが有用と考えられる。逆に、生成 AI モデルが試験問題に解答する、あるいはクイズの解答者として想定されている場合は、余計な説明はせずに答えのみを出力するのが正確・有用であると言える。いわゆるハルシネーションについても、文脈によっては有用であることもある。例えば、現実とは異なる世界を前提として異世界小説のアイデアを出すユースケースなどがある。また、生成 AI モデルを特定ドメイン・アプリケーションで利用する場合には、ドメイン固有の知識・文脈や、アプリケーションでの要求仕様によって、正確性・有用性の定義が変わるため、改めて正確性・有用性を定義する必要がある。

② 指示追従性・制御性

正確性・有用性に関連して、生成 AI モデルについて特に必要とされる要件として、回答が満たすべき条件や出力フォーマットなど人間の指示に忠実にしたがって生成を行うこと（指示追従性）、さらにシステム全体として人間が意図した形で動作させることができること（制御性）が挙げられる。前述したように、生成 AI モデルの出力は無制限であり事前に限定することができないため、指示追従性や制御性は特に重要な要件である。

③ 安全性

生成 AI モデルが備えるべきもう一つの重要な要件として、安全性がある。生成 AI モデルの安全性に対する脅威については３(1)節で詳述されているが、生成 AI モデルは統計的にコンテンツを生成しており倫理観を持っているわけではないため、生成 AI モデルが倫理に反したり危険なコンテンツを生成しない仕組みが必要である。安全性

の中にもいくつかのカテゴリがあり、犯罪や違法行為、ポルノに関するコンテンツを生成しないことは当然であるが、それ以外にも対処すべきコンテンツがある。例えば、NII の大規模言語モデル研究開発センターが開発している LLM の安全性のインストラクションデータ AnswerCarefully では、LLM が対処すべき安全性として５つのリスクタイプ、12 の有害カテゴリ（アダルト、ステレオタイプ・差別の助長、ヘイトスピーチ、メンタルヘルス、AI の擬人化、個人情報漏洩、組織・国家機密情報漏洩、違法行為への加担、非倫理的行為への加担、偽情報拡散への加担、誤情報による実被害、誤情報の拡散）を定義し、これらのカテゴリを網羅するデータセットを構築している[26]。５つのリスクタイプは、バイアス・差別・ヘイト・反公序良俗（非社会的な回答やアクション生成による脅威、名誉毀損・信用毀損、社会の価値観や文化の変化）、AI との対話によるリスク（高品質な生成メディアによる脅威）、情報漏洩（機密情報漏洩による脅威）、悪用（著作権侵害、ハッキングによる脅威、コード生成 AI に関する法的リスク）、誤情報（ハルシネーションによる脅威）と定義されており、本章で議論された脅威と課題の多くをカバーしている。

④　バイアスへの対処・公平性

生成 AI モデルは、多様な場面で利用されることを考えると、バイアスへの対処や公平性も必要である。生成 AI モデルが生成するコンテンツに差別や偏見が含まれないようにすべきであり、また、特定のグループに有利なコンテンツを出力することも避けるべきである。バイアスとしては、ジェンダー、人種、宗教、職業、国籍などについて、生成 AI が様々なバイアスを持つことが指摘されている[48]。ただし、差別や偏見は文化や時代によって変化し得るものであり、現在は差別・偏見と認識されていないものが将来的に差別・偏見とされる可能性がある。将来的に差別・偏見とされ得るものまで網羅して対処することは困難であるが、少なくとも現時点で明確に差別・偏見とされるものについて、不適切なコンテンツを生成しないよう対処することが求められる。

⑤　頑健性・セキュリティ

生成 AI モデルに関する要件の一つとして、従来の AI システムにおいても重要な要件であるが、システムの頑健性・セキュリティがある。攻撃者により AI システムが意図しない動作をさせられたり、システムを改変される危険性は排除すべきである。生成 AI モデルについては、生成内容を意図的に改変する（誤情報に基づくコンテンツを生成させるなど）といった攻撃が想定され、それに対する防御が必要となる。ただし、頑健性・セキュリティがどこまで求められるか、どのように担保すべきかは、アプリケーションや想定ユースケースなどによって変わり得る。また、生成 AI が出力するコンテンツを制御することは困難であることから、ユーザーが悪意を持たなくても個人

[26] https://llmc.nii.ac.jp/answercarefully-dataset/

情報や機密情報が漏洩する危険性がある。ガードレールを設けるなど、個人情報や機密情報が漏洩しないようにする対策も必要である。

⑥ 透明性

信頼される生成 AI モデルのためにもう一つ必要な要件として、透明性が挙げられる。透明性とは、生成 AI モデルの開発や動作について人間が理解できることを指す。生成 AI モデルは巨大なブラックボックスであり、現時点では入力から適切な出力が計算されるプロセスは完全には分かっていない。よって、生成 AI モデルの動作について完全な透明性を求めることは不可能であるが、様々な観点から生成 AI モデルの透明性を高めることが求められる。

まず、システムの透明性が挙げられる。生成 AI モデルの中心は巨大なパラメータであり、その動作を理解することはできないが、それを動作させるためのシステムは通常のプログラムである。どのようなアルゴリズムで学習されているのか、生成はどのように行われているのかなど、生成 AI モデルのプログラムの動作について利用者が理解できる必要がある。理想的にはソフトウェアをオープンにすることが考えられるが、オープンにできない場合でも、内部で使用しているアルゴリズムなどについては最大限情報提供することが求められる。また、生成 AI モデルは、大規模データを用いた機械学習によって構築されるため、学習データによってその性質が規定される。したがって、データの透明性、すなわち学習に用いたデータについて十分な情報提示が必要である。どのようなデータを学習に用いているかは、モデルの性質を理解するだけでなく、著作権など倫理面について問題がないか判断するためにも必要である。最後に、モデルの透明性、すなわち生成 AI モデルが入力から出力を生成するプロセスの理解について、これは現在の技術では未解決であるが、生成 AI モデルの動作を解明する研究開発を継続する必要がある。様々なテストで生成 AI モデルの挙動を分析したり、内部のパラメータを分析したりして仕組みを解明しようとする研究が進められている。

⑦ 説明可能性・説明責任

透明性に関連する重要な要件として、説明可能性がある。これは、AI モデルの出力について人間が理解できる形で何らかの説明がなされることを指す。生成 AI モデルにおいてよくある手法として、モデルの出力について生成 AI モデル自身に説明させたり、最終的な出力をするまでの推論プロセスを明示させたり、可視化させる手法が挙げられる。さらに、生成 AI モデルの開発者や利用者は、生成 AI モデルについて何らかの問題が起きた際の責任の所在を明確にし、十分な対応を行うことが求められる（説明責任）。上述のように生成 AI モデルの出力やユースケースは、あらかじめ限定することができないため、生成 AI モデルの出力や動作について事前に完全な保証をすることは困難である。しかし、何らかの問題が発生したときの責任の所在や対応プロセスがはっきりしていれば、ユーザーは安心して生成 AI モデルを利用することが

できる。

⑧ 資源効率

３(3)節で述べたように、現状、生成AIの開発・運用には、膨大な資源（計算リソース、データ、電力、資金）が必要である。特に電力消費は、環境負荷の面で問題視されつつある。また、基盤モデルのロボットなどへの搭載が進みつつあるが、エッジAIの応用ではリアルタイム応答や軽量実装のため、コンパクトで高速なAIモデルが必要になる。このようなニーズから、膨大な資源を要しない効率性の高いAIモデルが求められる。大規模な事前学習を行わない能動的な学習技術、蒸留などによるモデルのコンパクト化技術、処理の分散協調化技術、AI用の省エネルギーチップなどが研究開発されている。

⑨ 生成AIモデルが備えるべき要件のトレードオフ

以上、生成AIモデルが備えるべき要件を列挙したが、これらの要件にはトレードオフが存在する。例えば、正確性・有用性と安全性には明確なトレードオフがある。究極的に安全なシステムとは、何も出力しないものである。しかし、これは明らかに正確性・有用性がなく、意味がないシステムである。一方、正確性・有用性を最大化すると、犯罪行為に関する質問に対して詳細な情報を提供することになり、安全性が低下する。また、正確性・有用性と安全性のトレードオフの閾値は、ドメインやユースケースによって変わり得る。例えば、毒物の作り方について詳細に説明することは一般的には安全性が低い（危険性が高い）と言えるが、例えば化学物質に関する専門家が学術研究に利用するチャットシステムであれば、毒物を含めて詳細な作成方法を説明できる方が有用性は高いと言えるかもしれない。同様のことは他の要件についても当てはまり、例えば、完全な透明性を実現すれば、攻撃者に対して情報を提供することになり、安全性や頑健性が低下する。また、スケーリング則によるとモデルサイズや学習データを増大させると正確性・有用性を高めることができるが、これにより資源効率の問題は深刻化する。

⑩ 生成AIモデルが備えるべき要件を実現する方法

最後に、上記の要件を満たす生成AIモデルを実現するための方法について述べる。信頼される生成AIモデルとは、これまでの議論からすると人間あるいは社会の価値観・倫理観や要求に合うようにコンテンツを生成するものであると言える。このように、生成AIモデルが人間の価値観・倫理観や要求に合致するコンテンツを生成できることをアライメントという。生成AIモデルのアライメントを実現するための代表的な手法として、２(1)節で述べたファインチューニングと、評価がある。例えば、前述したインストラクションデータAnswerCarefullyを用いてファインチューニングを行うと、LLMは危険あるいは非倫理的なプロンプトに対して回答を拒否したり、その理由を説明することができるようになる。

ただし、ファインチューニングはいずれの手法もモデルのパラメータを調整するものであり、生成AIモデルがブラックボックスであり、出力を限定・制御できないことに変わりはないことに注意すべきである。つまり、上記の手法によっても、正確性・有用性や安全性を完全に満たすことは原理的に不可能である。よって、透明性や説明責任を担保することで、何らかの問題が起きた時に説明や対処ができる必要がある。さらに、生成AIモデルを利用する場面においては、生成AIモデル自体に頼るだけでなく、それを利用するアプリケーション・サービスやその運用において、正確性・有用性や安全性を確保する仕組みも必要となる。代表的なものとして、ガードレールが挙げられる。ガードレールは、生成AIモデルの入力や出力に対してフィルタリングなどのチェック処理を行う。どのような情報が有害であるかは文脈に応じて変化するため、入出力のチェック機能などは文脈に応じた判断が必要であり、このため比較的小規模なAIを用いる必要がある場合も多い。また、入力制御においては、より信頼性の高い関連情報を検索して生成AIモデルに入力したり、場合によっては生成AIモデルを用いるのではなく従前のルールベースの生成手法を活用したりすることで、生成の正確性・安全性を向上させることなども重要となる。

そして、生成AIモデルがどの程度アライメントを実現したか、人間の価値観・倫理観に合う出力ができるかを示すために評価が必要である。これまでたびたび議論しているように、生成AIモデルの出力は、あらかじめ限定することができず、ユースケースが限定されないことから、生成 AI モデルをどのように評価すべきかは自明ではない。これまで、英語を中心に正確性、有用性、安全性などを様々な観点で評価するベンチマークが数多く開発されており、これらを用いて評価を行い、その結果を開示することが求められる。さらに、特定ドメインやアプリケーションを定めることで変わる、あるいは新たに求められる要件がある。例えば、医療情報を提供するアプリケーションでは、法令遵守はもちろん、生成AIモデルの出力によって引き起こされ得る有害性・危険性を具体的に定めた上でそれらの評価・対策を行う必要がある。したがって、生成AIモデルの開発者だけでなく、生成AIモデルを利用したサービスやアプリケーションの開発者・運用者も、サービス・アプリケーションで要求すべき正確性・有用性・安全性やその他要件を定め、適切な評価を行う必要がある。

特に安全性に関しては、生成 AI モデルやそれを組み込んだアプリケーションに対して、十分に機能しているかどうかをテストする枠組みとして、レッドチーミングと呼ばれる攻撃シミュレーションが求められている。生成 AI モデルを市場投入する前にはレッドチーミングを実施することが、国際的な行動規範として G7 首脳間で合意されている[27]。これを実現するためには、攻撃用生成AIを活用した自動レッドチーミングなども重要になる。これにより、安全性・公平性・頑健性・セキュリティはもちろんのこと、多面的な観点から生成AIモデルが持つ潜在的な危険性を評価した上で、市場投入するかどうか判断することなどが求められるようになる。機械学習を組み込

[27] https://www.mofa.go.jp/mofaj/files/100573472.pdf

んだアプリケーションシステムについて、安全性・信頼性を確保して効率良く開発するための方法論・技術群を整備するソフトウェア工学的なアプローチは、機械学習工学や AI ソフトウェア工学と呼ばれる研究分野として、2018 年頃から活発に取り組まれるようになった[49][50][51]。この取り組みにおいて現在、生成 AI 対応の拡張が進められている。

最後に生成 AI のガバナンスについて述べる。２(3)節で生成 AI に対する規制の世界と日本の動きを概説した。本章でこれまで述べてきたように、生成 AI に関するリスクは多種多様であり、技術進歩も速いため、リスク状況は予測困難かつ急速に変化する。また、リスクへの最適な対応策は、ガードレール技術や組織のマネジメントシステムなどを複層的に積み重ねた複雑なものとなる。このような状況において、生成 AI の研究開発や社会での活用を積極的に進めていくためには、従来型の規制モデル、すなわち、法律があらかじめ細かな行為義務や禁止義務を設定し、規制対象者は決められた義務に従うという考え方を更新する必要がある。生成 AI の特徴を踏まえ、ガイドラインをはじめとするソフトローと、ハードローを柔軟に組み合わせた新たな制度設計を行うことが求められる。そこでは、関係するマルチステークホルダー間の透明性のある議論や、生成 AI の国際性を踏まえ国際的ルールメイキングへの参加・貢献も重要である。

4　生成 AI の活用による波及効果

(1)　科学技術の発展に対する効果

AI と科学技術の歴史は古く、既に AI の黎明期である 1960 年代には、知識ベースと推論ルールから構成されるエキスパートシステムが登場し、科学への適用が試みられていた。その後、計算機の普及とともに登場した情報検索、自然言語処理、データベース、機械学習、セマンティックウェブなどの情報・AI 技術は、科学のパラダイムシフトをもたらすと同時に、紙媒体からデジタルへの学術情報流通の変遷を支える基盤にもなった。

現在、科学の様々な領域において、モデルとデータをつなげ、仮説の生成と検証のサイクルを一体化するデータ駆動科学的な手法が注目されている[52]。また、そのための基盤として、研究データを含めた研究成果をオープン化して広く共有するオープンサイエンスが推進されている[53]。科学が対象とする問題が複雑化する中で、AI による科学の加速への期待は大きい[54]。実際、生成 AI の活用により、科学研究の各ステップは大きく進展している。

仮説の生成では、AI による高度なデータ解析が可能となり、バイオや材料科学に加えて、人文社会系でも効率的な仮説探索が行えるようになった。また、AI は文献解析を通じて新たな発想を支援し、実験計画の立案をサポートしている。仮説の検証においては、AI とシミュレーションを組み合わせた仮想世界での検証が進化しており[55]、物理世界では、ラボオートメーションによる実験の効率化が注目されている。論文による知識流通においては、AI が執筆支援を行い[56]、出版社も対応を表明している。また、自動査読が現実味を帯び[57]、要約や推薦、Q&A を通じたアクセス支援が強化されつつある[28]。

生成 AI の活用は、各学術ドメインで急速に進展しており、各国で様々なプログラムが推進されている。例えば、米国では「AI for Science Workshop」[29]が開催され、科学技術の各分野における AI の応用が議論されている。材料・物性科学分野では、AI が新素材の設計や物性予測を支援し、実験プロセスを効率化している。生命・医科学分野では、AI が遺伝子解析や薬剤設計において重要な役割を果たし、創薬や疾病予防の加速に貢献している。また、国立研究開発法人理化学研究所では、「AI for Science」のための TRIP プロジェクトが進められており、各学問分野における AI の活用が科学研究の新たな可能性を広げている。例えば、分子動力学シミュレーションと AI を統合し、タンパク質の構造や機能の予測を高精度かつ迅速に行うことで、新薬候補の探索や創薬プロセスの効率化を図っている[58]。

AI は、自然言語や画像・音声処理、情報検索、統計的機械学習、記号推論などの最先端の情報技術を結びつけて、研究者の日々の研究活動を効率化するとともに、科学のサイクルそのものを変容しようとしている。また、ネットワーク、セキュリティ、データ、大規模計算システムなどは、あらゆる学術領域を支える基盤としての役割を担っている。生成 AI の科学技術への適用においては、AI 技術の研究開発の推進や学術基盤の強化が

[28] https://www.nature.com/articles/d41586-024-02942-0
[29] https://ai4sciencecommunity.github.io/

前提として必要である。その上で両者が協調する枠組みを構築することで科学のパラダイムシフトが可能になる。そのような方向性を目指すものとして、「次世代計算基盤（富岳 NEXT）」におけるシミュレーションと AI の連携や、日本学術会議の「未来の学術振興構想（2023 年版）」におけるデータ基盤と利活用に関するグランドビジョンなどがある[59]。また、生命・医科学分野や材料・物性科学分野など戦略的に重要なドメインに対する科学研究向け生成 AI モデルの構築も構想されている[60]。さらに、知識の細分化・専門化による学問領域の分断化の問題が指摘される中で[61]、分野横断的な知の活用を可能とする知識基盤の構築に、生成 AI を活用することも重要である。

科学における AI の適用では、グランドチャレンジの設定が重要である。論文や研究データの量が増加の一途をたどる現状において、あいまいなノイズや情報の欠落などを含む膨大な情報の処理に、人間の研究者の能力が追い付かなくなっている[62]。AI で作業を代替することで、これまで不可能であった仮説空間の網羅的な探索が可能になる。このような考えに基づく AI のグランドチャレンジとして、Nobel Turing Challenge[63]がある。Nobel Turing Challenge は、AI による科学のパラダイムシフトを科学における産業革命になぞらえ、ノーベル賞に値する科学的発見を目指すものである。生命科学の分野で研究者による仮説の発見と探索を支援する AI を出発点とし、さらなる挑戦として、研究を自律的に行う AI エージェントの実現を掲げている。後者では、人間の研究者が仮説空間や実験環境をあらかじめ準備することなく、仮説の生成から検証を含む研究サイクルのすべてを AI が自動的に行うことで、研究サイクルの大幅な高速化や研究サイクルそのものの大規模な並列化が期待される[64]。

研究における生成 AI の活用に向けてのビジョン策定が進む一方で、その実現に向けては生成 AI の信頼性の担保が重要な課題となる。特に、研究の再現性のために生成 AI の透明性が求められており、研究データのバイアスや研究者の認知的限界が仮説空間を狭める可能性も指摘されている[62]。また、AI の役割が人間の支援に留まるのか、自律的に研究を進めるエージェントの実現を目指すのかについては、AI 規制の動向、研究倫理、安全性などを踏まえた議論が求められる。自律的な AI エージェントを目指す場合でも、研究のすべての過程を人間が理解できる形で言語化して共有するとともに、誰もが再利用できる知識へと蒸留するプロセスが必要であろう。このことは、科学における理解や正しさは何かという科学の根源的な問いに関わるものである。

最後に、生成 AI の研究利用が普及すると、個々の研究者の能力よりも、生成 AI の処理能力が研究の質や生産性に影響を与えると予想される。これまでは、我が国の学術振興に向けて研究者間の競争的な研究環境の形成が重視されてきたが、生成 AI の利活用の観点からは、我が国の学術コミュニティ全体としての研究力の強化が優先課題となる点に留意するべきである。

⑵ 産業分野への効果

① 産業面から見た生成 AI の価値

我が国では現在、労働力不足や長時間労働が社会課題となっており、働き方改革が

急務となっている。これまで長時間労働是正に向けた様々な取り組みがなされているものの、業務量自体の削減はあまり進んでおらず、根本的な解決が困難であった。生成AIは、多くの産業において業務効率化・業務量削減に資すると期待されており、労働に関する諸課題を解決する強力な一手となり得る。例えば、ビジネスパーソンの多くが、情報収集や資料作成、データ投入などに多くの時間を割かれており、長時間労働の一因となっている。2024年現在でも、生成AIを活用することで、Web上の情報を調べて見やすくまとめておくなど、作業の少なくとも一部を代替することができる。さらに、PCのデスクトップ全体を人と同じように認識し自動的に操作する試みも急速に進んでおり、近い将来、人と遜色のないレベルで、非定型的かつ複雑な作業の自動化が可能になると予想される。このように、生成AIの活用により、従来の常識とは全く異なるレベルでの業務効率化が実現できると考えられる。

また、業務の品質自体の改善にも、生成AIは貢献できる。これまで一部の人のみが持っていた特殊技能（多言語でのやりとり、プログラミング、動画作成、膨大な情報の処理など）を、あまねく人が活用できるようになる。現在でも、広告のコピーの原案を生成したり、ポスターのイメージ画像を生成するなどの活用が進められている。近い将来、特殊技能のコモディティ化がさらに進み、従来国内にのみ展開していたビジネスを容易にグローバルに展開できるようになったり、自社の複雑な業務を自動的に処理するプログラムを自前で作成・運用し、人手作業によるミスを抑制したり、より深く広い市場調査に基づいた経営判断をできるようになると期待される。

加えて、生成AIは、労働者の心理的な健康の維持にも貢献する。上述の業務効率化・業務量削減によって、より本質的な業務へ集中することができ、ストレスの軽減につながる。また、長時間労働を是正することで、睡眠時間や余暇時間の確保にもつながる。さらに、生成AIは従来のシステムと異なり、人との対話が可能である。すなわち、人の発言を直接受け止められるインターフェースとなることで、コールセンターのクレーム対応や、様々な現場での暴言への対応などを受け止める壁となり、そうした業務に従事する人々の心理的な健康を守ることが可能になる。

②　普及においての課題

生成AIの普及により、業務効率化や自動化が進むことで、多くの産業においてポジティブな変化が期待される一方で、既存のエコシステムに与える影響についても慎重に検討する必要がある。特に、オフィス業務やデジタル化された分野では、単純作業や中間業務が自動化され、一部の職種が消滅する可能性が高い。例えば、カスタマーサポートやデータ入力などの業務は、生成AIの導入によって効率化され、これまでそれらの業務に従事していた人々が職を失うリスクがある。また、法律文書作成やクリエイティブ業務の自動化により、士業（弁護士、司法書士など）やクリエイターも新たなスキルを習得する必要に迫られることが予想される。一方、生成AIは物理的作業を直接代替することはできない。物理的作業が可能なロボットも、複製が容易なデジタル領域ほど急速には普及が進まないと考えられる。そのため、物理的作業がコア

業務である産業、例えば一次産業、二次産業、医療・福祉の分野では、生成AIによる自動化の影響が相対的に小さく、市場における労働価値が高まると予想される。このように、生成AIの普及により、労働市場全体の再編が進むと考えられる。それに並行して、生成AIを活用したリスキリングもより盛んになるであろう。これまで以上に、社会全体での就労支援や再教育の取り組みが重要になると考えられる。

また、生成 AI の普及には膨大なデータが必要であり、そのデータの収集や使用に関しても課題がある。特に、生成AIが誤ったデータや低品質なデータを基に動作する場合、３⑵節や３⑷節で述べたような誤情報やバイアスを含むアウトプットが生成されるリスクがある。例えば、医療や金融といったデータの正確性が非常に重要な分野では、不正確なデータを基にAIが判断を下すことで、深刻な社会的影響をもたらす可能性がある。このため、生成AIが利用するデータの信頼性を確保し、質の高いデータ管理プロセスを導入するとともに、堅牢性の高いチェック機構を併用することが重要である。また、個人情報やプライバシーに関わるデータの取り扱いについても、適切な規制と利用者の同意を得るための明確な手続きが必要となる。

さらに、生成 AI の公平性や安全性を確保するための対策も重要である。あまねくデータには常に何らかの偏りが存在するため、それを学習したAIの出力にも、学習データに応じたバイアスが生じ、公平性が損なわれるリスクがある。例えば、採用や融資判断などの場面で、偏ったデータに基づいたAIの判断が、不公正な結果を招く危険性がある。学習データの偏りを完全に除去することは原理的に困難であるため、単に偏りを減らす対策では不十分である。AIの判断の説明性を高める取り組みや、AIに公平な観点を学習させるなどのAIの改良を進めるとともに、我々人間がAIを利用する際に、AIの出力にバイアスが生じ得ることを認識し、AIの判断を鵜呑みにしないことが重要である。また、生成AIの安全性を確保するためには、システムが悪用されないようにするためのセキュリティ対策や、AIが人命に関わる領域で誤った判断を下さないような規制の強化も必要である。生成 AI の信頼性と安全性を高めビジネスを円滑に進めるためには、３⑷節に挙げた技術的な対策とともに、法制度や倫理的ガイドラインの整備、第三者機関による認証制度の確立が急務である。

③　各種権利処理とデータ流通

生成 AI の開発・運用には、大量の学習データや RAG のための知識ベースが欠かせないが、これらのデータを商材とするデータホルダー（特にクリエイター）への対価を支払わずに、生成 AI によって同等のコンテンツや二次創作物を生成できる場合には、データホルダー側と生成AI開発者側との利害が相反することがある。一方で、生成 AI 事業者の収益の一部がクリエイターに使用料などで配分されるなどのエコシステムを構築できれば、クリエイターの創作意欲を刺激しデータを基軸とした産業振興にもつながる。既に動画プラットフォームなどで実現されているエコシステムを、生成AI向けに拡張しつつ生成AIで得られた莫大な収益を配分することで、クリエイターにとってより高い収益が望めるプラットフォームにしていくことなどが望まれる。

一方で、各種権利を無視して生成AIのモデルを学習・利用する行為を統制するべく、何らかの規制と判定技術が必要になるであろう。さらに、生成物が各種権利を侵害する可能性を測るべく、生成物と権利保護されたコンテンツを比較・照合する技術も必要となるかもしれない。

④　生成AIによる生成物の明示とフェイクコンテンツの防止

AI生成物が実物と見分けがつかなくなることにより、情報操作などにつながる危険性がある。また、AIの助けを借りた達成なのか、個人の実力としての達成なのかの判定も難しくなり、公平性の観点でも社会問題になる可能性がある。このためこれらを判別するための技術も必要となる。これらが担保されて初めて社会に認知されるツールとなる。この技術は未だ研究段階であり、今後の進展が望まれる。

⑶　社会的な波及効果

生成AIの効果は、科学・産業界に留まらず、我々の日常的な社会活動にも波及している。もはや社会的インフラと言ってよい情報検索についても、ウェブ検索が生成AIに代替されつつある。ウェブ検索の機能は、知りたい情報へのリンク群の提示であり、ユーザーは、それらの中から適切な情報を探し出す必要がある。これ対し、生成AIは知りたい情報そのものを直接回答してくれるため、手間が大幅に低減される。再三述べたように、生成AIによる回答には誤情報が含まれるリスクがあり、また生成AIの結果だけをみて、その結果の背景となる一次情報を見なくなるおそれがある。しかし、RAGによってウェブ検索結果を引用しつつ回答を生成する技術も進化しており、生成AIから情報検索を利用することや、生成AIの結果に一次情報への参照を加えるなどの対策も進んでおり、生成AIの利用は今後も増加するだろう。

コミュニケーションバリアを取り除くにも生成AIが活用されている。その顕著な例が翻訳であり、入力された言葉を別の言語に変換した結果を得る際に生成AIが利用されている。最近のリモート会議システムやプレゼンテーションソフトウェアには翻訳機能が搭載されており、リアルタイムでの異言語間コミュニケーションを可能にしている。他言語と大きく異なる言語体系を有する我が国にとって、この翻訳によるバリア解消は、生成AIがもたらす大きな恩恵の一つである。また、長い文章や一般には難解な文章を分かりやすく要約してくれたり、自身の文章が、他者により伝わるように推敲してくれたりすることも、コミュニケーションバリア解消に寄与する生成AIの機能である。

生成AIの創造性も、様々な形態で日常生活に利用されつつある。創造的な文章生成については、挨拶文やメールの生成支援、料理レシピや旅行計画の提案、投資アドバイスなど多岐に渡った利用が考えられる。画像や音の生成についても、アートやデザインの生成や音楽の生成など、趣味レベルから専門家レベルまで、利活用が進んでいる。こうした生成結果をどう評価し受け入れるかは、人間に委ねられているものの、創造支援としての生成AIの役割は今後ますます増大するものと予想される。

生成AIによる社会への最大の波及効果として、教育への影響が挙げられる。現在の大

学教育では、科学技術の高度化により学生・大学院生に教えるべき内容が増加するとともに、学生・大学院生の知識や関心事も多様化しており、従来の画一的な教育には限界が生じている。また、大学の研究室内でも学生同士が議論することが難しくなりつつある。その中で生成 AI、特に対話 AI は、周囲に議論の相手がいない場合であっても、一定の専門性のある議論を行う機会を作り得る。一般的な授業においても、学生が抱える疑問や不明点について手軽に生成 AI に質問することで回答が得られれば、教員に質問せずとも授業の理解を深められる。また、効率的な教材作成、リアルタイムフィードバック、多言語対応、創造的な学習活動への寄与など、教育現場おいて生成 AI は様々に利用し得る。

一方、生成 AI の教育応用については、様々な危惧も存在する。UNESCO は、生成 AI の急速な発展が国家規制の進展を上回る中、教育機関の対応が追いついていない現状に強い危機感を抱き、2023 年に教育および研究分野における生成 AI 活用に関する人間中心のアプローチを提唱する提案を発表した[65]。この中では、人間の主体性、包摂性、公平性、男女平等、言語的･文化的多様性、多様な意見や表現を促進する中核的な人間的価値観に対して 生成 AI が及ぼす可能性のある潜在的なリスクの評価を示している。さらに、データプライバシーの保護を義務付け、使用年齢制限を考慮するなど、政府機関が生成 AI ツールの使用を規制するための重要な手順を提案している。特に、初等中等教育へのデジタル技術の導入にはスウェーデンのように慎重な国もある[30]。

我が国でも、初等中等教育段階における生成 AI の利用に関する暫定的なガイドラインが、文部科学省により 2023 年７月４日に発行された。このガイドラインは、生成 AI の高いポテンシャルを有効活用することの促進と、プライバシーや著作権の侵害といった利用上のリスクや不正使用に対する注意喚起という二つの側面を持っている。さらに同省は 2024 年 12 月 26 日、ガイドライン（指針）の第２版を公表した[66]。その中で生成 AI の利用には情報モラル教育が欠かせないとして、「特に小学校段階の児童に直接使用させることには慎重な対応を取る必要がある」と明記している。また、我が国の大学においても、様々な利用ポリシーが定められつつある。

教育は国家の根幹をなすものであり、生成 AI の限界やリスクを把握した上での慎重な議論が必要である。D.Long らは、AI 技術を批判的に評価しながら、効果的に対話し、共創し､家庭や職場で AI をツールとして使用するための 17 個の能力(コンピテンシー)を挙げている[67]。例えば、AI を使用している技術とそうでない技術を区別する能力、AI を開発するための多様な方法を認識しAI を使用する技術を特定する能力などである。そして、それらの能力に対する授業や教材、ツールの設計上の考慮点について述べており、我が国の教育における生成 AI の利活用および AI リテラシー教育のポリシーを考える上で大いに参考となる。

[30] https://newsphere.jp/national/20231004-1/

５　提言

これまで述べてきた通り、生成 AI の急速な進展には、あらゆる学術分野、産業分野、そして社会全体に大きな影響を持つという包括性、将来的には人間と共存する知的レベルとなり得る革新性、さらに、それが急速に加速度的に進展するという加速性などの特徴がある。それゆえに、脅威や課題が存在するとともに、社会への大きな波及効果があり、人類社会の重要課題に対して解決策を提供する可能性がある。

我が国は、長い歴史の中で、外国の制度や技術を柔軟に取り入れ、それを発展・改良して世界に還元するということを行ってきた。また、AI・人工知能やロボットに親近感を持ち、共存することに比較的抵抗感の少ない国民性を有する[68][69][70]。

生成 AI の世界的進展が留まる気配のない中で、我が国は、リスク対策についても十分に工夫をしながら、生成 AI の研究開発や社会での活用を積極的に進め、人類と AI の共存社会のデザインで世界をリードすべきである。この基本的考え方のもとに、生成 AI を受容・活用する社会の実現に向けて４つの観点から計 12 の具体的な提言を行う。

(1)　生成 AI 研究開発の望ましい体制

①　生成 AI の技術開発を国家戦略として位置づける

生成 AI は広範な科学技術分野、産業分野に波及し、我が国の国際競争力を左右する重要技術である。生成 AI の研究開発は、ビッグサイエンス化し、多くの力を結集して取り組むことが必要になっている[71]。日本の技術競争力を強化するため、国家戦略として開発を推進するべきである。その研究開発のための予算の確保、および研究体制の整備を図る必要がある。

また、生成 AI の知識の包括性を活かし、様々な学術分野の横断的連携、さらに、各産業分野（製造業、医療、教育、公共サービスなど）と学・官の連携を推進すべきである。そこでは生成 AI の技術活用の情報・ノウハウの共有が重要であることから、オープンな研究開発の取り組みへの支援を重視・強化することが必要である。

②　生成 AI の研究基盤の強化と国際的研究連携の推進

基盤から応用にまたがる広範囲の生成 AI 研究コミュニティを強化し、データセット、計算リソース、研究費などの研究資源を拡充する。特に、生成 AI の開発には膨大なデータが必要であるため、プライバシーやセキュリティに配慮したデータインフラの構築を支援するとともに、公共データの開放や産業界とのデータ共有プロジェクトを奨励することが必要である。

さらに、国際的な研究連携を通じた知見共有や人材交流を加速させる必要がある。これによって、先進的な知見や最先端の技術を相互に活用し、地球規模の課題解決や科学技術のさらなる発展につながることが期待される。

③　生成 AI 開発における透明性の確保と AI ガバナンスへの包括的な取り組み

AI の開発において、作られた AI による判断や行動が人間の価値観や倫理観に合うことが極めて重要である。AI がどのように判断し行動をとるかは、学習時に使われた学習データや学習手法（特に指示学習やアライメント）によって規定される。そのため、学習データや学習手法を含む開発プロセスについて透明性を確保することが重要である。

さらに、たとえ正しく AI の学習目標を設定していたとしても、AI が自ら設定する副目標が、意図せずに人間の価値観や倫理観と矛盾する場合がある。このような事態を避けるために、AI の設計・開発・評価においてガイドラインを作成してリスクを最小化するとともに、こうした問題についての理解を深めるための研究およびこれらを回避するための技術開発について共同の作業が必要である。

また、上記の観点を含む AI ガバナンスの国際的ルールメイキングに、日本としての考え方を反映させる活動も重要である[72]。

⑵　生成 AI モデルの適切な運用

①　生成 AI に対する攻撃を検知・回避する頑健なシステム構築

生成 AI が今後、社会で重要な役割を果たしていく中で、これらのモデルはサイバー攻撃や物理的攻撃から適切に保護される必要がある。具体的には、敵対的入力に対する頑健性強化や攻撃の検知に関して、ベンチマーク作成・評価を含む研究を進め、モデル出力が一定の安全性を保つことが保証される仕組みを構築する必要がある。

また、組み込み機器や IoT デバイスに生成 AI が搭載される場合、物理的にデバイスが奪われたり、悪意のある改竄を受けたりする可能性が高まる。こうした攻撃を検知・回避する必要がある。特に今後、生成 AI が自動運転車やロボットなどの判断や制御に使われる機会が増えてくる。モデル開発者やシステム提供者は、モデル開発過程から提供時（ファインチューニングを含め）までにモデルが改竄されていないことを証明する仕組みを導入する必要がある。また、システムを複数の生成 AI から構成し、一部が改竄されている場合においても全体としては動作が保証される仕組みを導入するなど、これまで以上に頑健なシステムを作ることが求められる。

②　AI 利用のリスク最小化と迅速に問題に対処する体制の整備

AI を利用、運用する際にもたらされるリスクを最小化するための体制構築が必要である。今後も急速に発展する AI 技術の進展に柔軟に対応できるとともに、問題が発生した場合に迅速かつ適切に対処できる体制を整えることが必要である。

AI モデルを利用、運用する際にはその開発会社・組織、提供会社と協力した上で、その技術的特性や倫理的影響についてあらかじめ調査することが求められる。特に、海外の AI モデルを利用する場合には国内法が及ばないため、国際的な協力を通じて AI 技術の標準化やベストプラクティスを共有し、グローバルな視点での AI の発展と運用を推進することが重要である。

③　人間中心の原則に基づく持続可能な社会の実現に向けたAI利活用の継続的議論

AIシステムの設計と運用においては、科学的・技術的な問題だけでなく、価値観や倫理観、さらには社会全体への影響を考慮することが必要である。AIが多くの仕事をサポートし、あるいは置き換える可能性が高いことから、その恩恵が公平に配分されることが求められる。地球規模の課題や社会・経済にとって最重要な問題、例えば環境問題の解決や社会的な格差の是正に向けたAIの利活用・運用を優先すべきである。市場原理や競争原理に任せるのではなく、他方で硬直的な規制によるイノベーションへの弊害も踏まえた上で、人間中心の原則に基づく持続可能な社会の実現に向けて適切なインセンティブ設計や規制設計（規制緩和を含む）を通じてAI技術を適切に活用するための方策を継続的に検討するべきである。

⑶　責任ある生成AI実装に向けた制度設計

①　アジャイルかつマルチステークホルダー型のガバナンスの志向

人間中心の原則に基づく持続可能な社会の実現に向けた、責任ある生成AIシステムの設計・運用を促すためには、国家や一部の団体がルールを一律に決定するトップダウン型のガバナンスではなく、各事業主体が、バリューチェーン上での位置づけ（AI開発者、AI提供者、AI利用者など）や、ステークホルダーに生じるリスクの性質や量を踏まえて、最適な価値のバランスおよびその達成手段の設計・運用を行う、アジャイル（迅速・反復的）かつマルチステークホルダー型のガバナンスを志向すべきである。そして、そのような新たなガバナンスを実装するためには、以下に示すように、制度設計に関する官民の役割の変化が求められる。

②　政府の役割：オープンなルール形成・情報共有の促進、制裁に関する新たな制度設計

政府は、トップダウンで閉鎖的なルール形成を行うのではなく、政策影響評価の初期段階から議論を公にするとともに、広く一般からも意見を募るなど、オープンなルール形成を促進すべきである。また、事業者が主体的に責任あるAIを実装することを後押しするため、リスクマネジメントやステークホルダーコミュニケーションに関する枠組みを提供したり、法令の解釈に関するガイダンスを提供したり、事例に応じたベストプラクティスを共有したりするなど、ガバナンスに必要なツールの提供や法的な予見可能性を高めるための取り組みを推進すべきである。

仮に新たな規制を行う際には、イノベーションを阻害しないため、対象となるリスクが規制を必要とするだけの重大なリスクであるか（リスクベースの視点）、および規制がAIのみを差別的に重く規制する内容となっていないか（技術中立性の視点）を慎重に検討すべきである。

法的制裁については、複雑な要因が絡み合うリスクについて、すべての責任を特定の人物や組織に負わせ糾弾するような方法ではなく、事故の全体像の理解と将来に向けたシステムの改善を目的とし、関係者による事故調査への積極的な協力を促すよう

な制度設計を検討すべきである。例えば、刑事責任や課徴金の判断において、一定の標準への適合性が確認された場合には制裁の対象としない取り扱いや、事故後に情報を提出し原因究明に協力した事業者については制裁を免除・減軽するなどの取り扱いとすることも検討に値する。また、被害を受けた個人が、迅速かつ低コストで補償を受けられるような救済制度の整備も重要となる。

③ 民間主体の役割：主体的なリスク評価と AI ベネフィットの最大化、ガバナンスの恒常的な改善

民間主体による生成 AI のリスクマネジメントにあたっては、規制やガイドラインなどに記載されていることしか行わないという受け身の姿勢や、あるいは少しでも法令の適用関係が不明確なグレーゾーンには踏み込まないという消極的な姿勢ではなく、主体的にリスク評価を行い、そのリスクを社会にとって受容可能な範囲に収めつつ、AI がもたらす正のインパクトを最大化するような技術・プロセスおよび組織の設計・運用を行うべきである。また、そのような取り組みについて、政府と市民を含むステークホルダーに対する十分な質と量の情報開示を行い、アカウンタビリティを尽くすとともに、ステークホルダーからのフィードバックを得て、常にガバナンスのあり方を改善できるような体制を整備する必要がある。経営層は、そのような体制整備について責任をもってコミットすることが求められる。

(4) 生成 AI モデル以降の教育とリテラシー

① AI との共存・共生のための社会変革に対応する人材育成

生成 AI をはじめとする技術変革は社会構造を大きく変えていく可能性があり、これを理解・活用しないことは国家的リスクである。そのため、社会全体での教育やリスキリングに取り組んでいくことが必要であり、それを推進するためのリテラシーを持つ人材の養成と教育プログラムの推進、リスキリング支援があまねく必要である。また、生成 AI の研究開発に取り組む高度人材育成も同時に行う必要があり、特に AI、HPC、ネットワークなど、基盤と AI をつなぐフルスタック人材の育成が急務である。また、内閣府などが指摘するようにこれらの人材は都市部に集中しており[73]、さらに、スタートアップや AI 研究大学も都市部に集中しているため[74]、地方での人材育成が進まず、そのために、地域格差をますます助長する可能性がある。地域医療格差と同様の問題が、今度は社会全体に広まる可能性がある。AI による改革を進める人材育成において、地域格差に配慮し、むしろ地域格差を解消することを目指すべきである。

② AI との共存を目指した新たな教育への転換

AI ひいては ICT を利用した教育をどの段階でどのように導入するかについては、様々な議論がある。特に、AI の限界、AI の持つ様々なリスク、AI への強度依存、カ

ンニング、学生と教員の関係の希薄化、教育の質の低下への懸念、従来の評価方法の限界を考慮しつつ、慎重な議論が必要である。その上で、AI の活用を前提として AI との共存を目指した新たな教育への転換を図るべきであり、小中高から大学に至るまで、教育のあり方の見直しを行うべきである。

そのためには、AI の利活用に対して、現状のいささか防御的なポリシーを改め、AI の積極的な利活用のための我が国としてのポリシーを定める必要がある。そのポリシーの中では、AI の能力や仕組みを理解し、その出力や行動に対する利点や限界を踏まえる必要がある。その上で、従来の知識の伝達に偏重するのではなく、AI を批判的に利用し、課題を解決し、創造する能力を高める教育・カリキュラムが必要である[75]。また、こうしたカリキュラムの設計には、AI の利用や開発において、倫理的な問題を常に考慮し、人権や社会の利益を重視する ELSI の側面を必ず含むことが必要である[39]。このような議論を市民を含めた広いステークホルダーとの対話の上で慎重に進めることが望まれる。

さらに、AI を含む ICT 技術は、このような議論を踏まえて急速に進化することが予想されるため、先進的な事例を共有しながら、常にこのような議論を続け、カリキュラムやポリシーをリビングドキュメントとして改定していく体制が必要である。大学共同利用機関法人情報・システム研究機構および NII が主催する「教育機関 DX シンポ」では、AI 活用教育のあり方、議論を基盤にした学習、PBL (Project Based Learning) を拡張した学習方法などが議論されており、新しい教育について情報共有・議論する場となっている。このような活動を国として支援することも重要である[76]。

③ AI の学際性を活用するための学術分野間および産学間の対話・連携の促進

AI の活用は学際性・包括性を有する。すなわち、ありとあらゆる学問を網羅した知識を集積したものから答えを導き出し、学術を学際的に深めることに資する。このことは、複合的な社会課題の解決につながるが、一方で、AI の回答を批判的に吟味する能力を科学者も持ち合わせなければならないことを意味する。科学者が高い AI リテラシーを身につけることはもとより、学術分野間の対話や連携をさらに深めていく必要がある。また、このことは産業界も同様である。人類の長期的な発展につながる立場からの議論を、産官学を含めた広いステークホルダーで深めていくことに努力を払うべきである。

＜用語の説明＞

ChatGPT	OpenAI 社がユーザーとの対話的なサービスを目的として2022 年 11 月に公開した大規模言語モデルに基づく対話システム。
Common Crawl	インターネットをクロールして得られた Web データを公開する非営利団体の名称であるが、公開されているアーカイブデータを指すことも多い。
ELSI	Ethical, Legal and Social Issues/ Implications の略で、新たな科学技術の発展に伴って生じる倫理的、法的、社会的課題のこと。
GIGA スクール	文部科学省が 2019 年に開始した、全国の児童・生徒に一人１台の端末と高速大容量の通信ネットワークを含む教育 ICT 環境を整備するための取り組み。
Llama	Meta 社が 2023 年３月より公開している一連の大規模言語モデルの名称。2024 年 12 月現在では、最新となる Llama3.3 がソースコードも含めて公開されている。
記号推論	論理式や数式のように記号的に表現された推論規則によって論理的帰結を導き出す手法。
強化学習	初期状態から目的とする状態に遷移する知的な行動をモデル化した機械学習手法。行動には報酬が伴い、その報酬を最大化することで目的とする状態に到達できるように学習を行う。
教師あり学習	人手などにより正解のラベルを付与した学習データを用いた機械学習手法。
ジェイルブレイク	ソフトウェアやデバイスの脆弱性をついて本来制限された機能や権限を使えるようにすること。
自己教師あり学習	正解のラベルが付与されない学習データに対して、自動的に正解ラベルを推測して人手などの介在なしに教師あり学習を実行する機械学習手法。
自己注意機構	入力されたデータのうちどの部分に注目して出力を生成するかを計算する仕組み。
指示チューニング	ユーザーが生成 AI モデルに対してどのような出力を求めているのかを表現した入力（これを指示と呼ぶ）と期待される出力とを学習データとしてモデルをチューニングすること。
指示追従性	大規模言語モデルがユーザーの指示に対して、その意図に沿った出力を生成できる能力のこと。
蒸留	大規模で性能のよいモデルを教師モデルとし、より小規模なモデルを生徒モデルとして、入力データと教師モデルの出力をもとに生徒モデルを学習させる手法。
深層予測学習	ロボティクス分野で外部環境に働きかける入力や感覚を変化させ、外部世界や自身に関する予測と、実際に得られる感覚との誤差を修正しながら環境を理解しそれに適応してゆく深層学習手法。

生成 AI	様々なコンテンツを生成できる AI 技術やそれを用いたシステムの総称であるが、明確な定義は見当たらない（２(1)節を参照のこと）。
潜在変数	統計学や機械学習において、観測されるデータに影響を与えるが直接観測できない変数のことを潜在変数と呼び、統計的モデルにおける特徴分析やデータ圧縮に用いられる。
ソフトロー	民間で自主的に定められているガイドラインのほか、行政府が示す法解釈なども含む広い概念。法的な拘束力のある法律・条例などを指すハードローと対比する用語として使われる。
大規模言語モデル（LLM）	もともと言語モデルとは自然言語処理において単語列の出現確率を予測するモデルを意味したが、その後大規模なニューラルネットワークと学習データを用いて様々なタスクを実行できる汎用モデルが構築されるようになり、これを大規模言語モデル（Large Language Model：LLM）と呼ぶようになった。
知識ベース	エキスパートシステムなどのルールベースにより推論を行う際に利用できる事実や前提知識をデータベース化したもの。人間にとって可読な形式のものもある。
敵対的生成ネットワーク	データを生成する生成器と、与えられたデータが人工データかどうかを判別する識別器を組み合わせ、相互に学習を行うことで実データと区別できないほど極めて自然なデータを生成する生成器を構築できるようにしたニューラルネットワークシステム。
敵対的入力	対象となるシステムに意図的に誤った出力を生成させたり、予測を誤らせるような攻撃的な入力のこと。
パブリシティ権	有名人や著名人が、自己の氏名や肖像などが商品の販売などを促進する顧客吸引力を有する場合、対価を得て第三者に排他的に使用させることができる権利（３(2)節③を参照のこと）。
ファインチューニング	大規模データによって事前学習された生成 AI のベースモデルに対し、正確性・有用性や安全性を満たしたコンテンツを生成する、あるいは特定のタスクやドメインに適応させる目的で、モデルを調整する手法（２(1)節を参照のこと）。
プロンプトインジェクション	AI システムに与える指示（プロンプト）を悪用して、本来出力が抑止されているような情報を意図的に生成させる攻撃手段。
べき乗則（スケーリング則）	モデルとデータを単純に増大させることで、性能がべき乗則にしたがって向上することを経験的に示したもの（３(3)節を参照のこと）。
変分オートエンコーダ	観測されるデータは、いくつかの観測不可能な変数（これを潜在変数と呼ぶ）の確率分布に基づいて生成されるという確率的なモデル化をしたもの。入力データをこれらの変数で表

	現し、さらに出力として元データを再構築することで学習を行う。
レッドチーミング	セキュリティ対策の有無や有効性をテストするために、対象となる組織やシステムの脆弱性などを検証する敵対的な攻撃手法のこと。
ロボット基盤モデル	ロボット向けに多様なタスクを実行できるように学習された汎用モデル。

＜参考文献＞

[1] 文部科学省初等中等教育局.「初等中等教育段階における生成 AI の利用に関する暫定的なガイドライン」, https://www.mext.go.jp/content/20230710-mxt_shuukyo02-000030823_003.pdf, 2023

[2] 広島 AI プロセス, https://www.soumu.go.jp/hiroshimaaiprocess/, 2023

[3] G サイエンス学術会議 2024 共同声明, https://www.scj.go.jp/ja/int/g8/, 2024

[4] サイエンス 20 (S20) 2024, https://www.scj.go.jp/ja/int/s20/index.html, 2024

[5] Rishi Bommasani et al. “On the Opportunities and Risks of Foundation Models,” arXiv:2108.07258, 2021

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeffrey Dean. “Distributed Representations of Words and Phrases and their Compositionality,” Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, pp. 3111-3119, 2013

[7] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Proceedings of the 3rd International Conference on Learning Representations (ICLR-2015), pp. 1-14, 2015

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. “Deep Residual Learning for Image Recognition,” Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2016), pp. 770-778, 2016

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. “Generative Adversarial Nets,” Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems (NIPS-2014), pp. 2672-2680, 2014

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. “Improving Language Understanding by Generative Pre-Training,” OpenAI blog post, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. “Attention Is All You Need,” Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NIPS-2017), pp. 5998-6008, 2017

[12] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. “Language Models are Unsupervised Multitask Learners,” OpenAI blog post, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, 2019

[13] Tom B. Brown et al. “Language Models are Few-Shot Learners,” Advances in

Neural Information Processing Systems 33 (NeurIPS 2020), Vol.33, pp.1877-1901, 2020

[14]OpenAI. “GPT-4 Technical Report,” arXiv:2303.08774, 2023

[15]Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, Furu Wei. “Retentive Network: A Successor to Transformer for Large Language Models,” arXiv preprint arXiv:2307.08621, 2023

[16]Albert Gu and Tri Dao. “Mamba: Linear-Time Sequence Modeling with Selective State Spaces,” arXiv preprint arXiv:2312.00752, 2023

[17]Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, Shanshan Li. “At Which Training Stage Does Code Data Help LLMs Reasoning?” Proceedings of 12th International Conference on Learning Representations (ICLR-2024), arXiv:2309.16298, 2024

[18]Patrick Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” arXiv:2005.11401, 2020

[19]Diederik P. Kingma and Max Welling. “Auto-Encoding Variational Bayes,” Proceedings of 2nd International Conference on Learning Representations (ICLR-2014), arXiv:1312.6114, 2014

[20]Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer. “High-Resolution Image Synthesis with Latent Diffusion Models,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2022), pp. 10684-10695, 2022

[21]Lvmin Zhang, Anyi Rao, Maneesh Agrawala. “Adding Conditional Control to Text-to-Image Diffusion Models,” Proceedings of International Conference on Computer Vision (ICCV-2023), pp. 3813-3824, 2023

[22]William Peebles and Saining Xie. “Scalable Diffusion Models with Transformers,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV-2023), pp. 4172-4182, 2023

[23]Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision,” Proceedings of the 38th International Conference on Machine Learning (ICML-2021), pp. 8748-8763, 2021

[24]Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee. “Visual Instruction Tuning,” arXiv:2304.08485, 2023

[25]Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, Abdelrahman Mohamed. “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 29, pp. 3451-3460, 2021

[26]Alexandre Défossez, Jade Copet, Gabriel Synnaeve, Yossi Adi. “High Fidelity

Neural Audio Compression," Transactions on Machine Learning Research, arXiv:2210.13438, 2022
[27]Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, Marco Tagliasacchi. "SoundStream: An End-to-End Neural Audio Codec," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 30, pp. 495-507, 2022
[28]Zalán Borsos et al. "AudioLM: A Language Modeling Approach to Audio Generation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 31, pp. 2523-2533, 2023
[29]Felix Kreuk et al. "AudioGen: Textually Guided Audio Generation," The Eleventh International Conference on Learning Representations (ICLR), arXiv:2209.15352, 2023
[30]Andrea Agostinelli et al. "MusicLM: Generating Music From Text," arXiv:2301.11325, 2023
[31]Jade Copet et al. "Simple and Controllable Music Generation," Advances in Neural Information Processing Systems 36 (NeurIPS 2023), pp.47704-47720, 2023
[32]Michael Ahn et al. "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," arXiv preprint arXiv:2204.01691, 2022
[33]Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, Marco Hutter. "Learning Robust Perceptive Locomotion for Quadrupedal Robots in the Wild," Science Robotics, Vol. 7, Issue 62, doi:10.1126/scirobotics.abk2822, 2022
[34]Open X-Embodiment Collaboration. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6892-6903, 2024
[35]Hiroshi Ito, Kenjiro Yamamoto, Hiroki Mori, Tetsuya Ogata. "Efficient Multitask Learning With an Embodied Predictive Model for Door Opening and Entry With Whole-body Control," Science Robotics, Vol. 7, Issue 65, doi:10.1126/scirobotics.aax8177, 2022
[36]Zipeng Fu, Tony Z. Zhao, Chelsea Finn. "Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation," arXiv:2401.02117, 2024
[37]Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, Shuran Song. "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," arXiv preprint arXiv:2303.04137, 2023
[38]AI 事業者ガイドライン（第 1.0 版）, https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240419_report.html, 2024
[39]カテライ アメリア・井出和希・岸本充生.「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」, 大阪大学 ELSI NOTE, https://doi.org/10.18910/90926, 2023

[40]国立研究開発法人科学技術振興機構 研究開発戦略センター.「人工知能研究の新潮流2～基盤モデル・生成AIのインパクト～」, 戦略プロポーザル, CRDS-FY2023-RR-02, https://www.jst.go.jp/crds/report/CRDS-FY2023-RR-02.html, 2023
[41]Martin Treiber. “The Secrets of GPT-4 Leaked?” IKANGAI, https://www.ikangai.com/the-secrets-of-gpt-4-leaked/, 2023
[42]Nestor Maslej et al. “The AI Index 2024 Annual Report,” AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, https://aiindex.stanford.edu/report/, 2024
[43]Epoch AI. “Data on Notable AI Models,” https://epochai.org/data/notable-ai-models, 2024
[44]Microsoft. “OpenAI Forms Exclusive Computing Partnership With Microsoft to Build New Azure AI Supercomputing Technologies,” Microsoft News Center, https://news.microsoft.com/2019/07/22/openai-forms-exclusive-computing-partnership-with-microsoft-to-build-new-azure-ai-supercomputing-technologies/, 2019
[45]Microsoft. “Microsoft and OpenAI Extend Partnership,” Microsoft Corporate Blogs, https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/, 2023
[46]Jared Kaplan et al. “Scaling Laws for Neural Language Models,” arXiv:2001.08361, 2020
[47]Niklas Muennighoff et al. “OLMoE: Open Mixture-of-Experts Language Models,” arXiv:2409.02060, 2024
[48]Isabel O. Gallegos et al. “Bias and Fairness in Large Language Models: A Survey,” Computational Linguistics, 50(3), pp. 1097-1179, 2024
[49]石川冬樹・丸山宏 (編) .「機械学習工学」, 講談社, 2022
[50]国立研究開発法人科学技術振興機構 研究開発戦略センター.「AI応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立」, 戦略プロポーザル, CRDS-FY2018-SP-03, https://www.jst.go.jp/crds/report/CRDS-FY2018-SP-03.html, 2018
[51]中島震.「AIリスク・マネジメント: 信頼できる機械学習ソフトウェアへの工学的方法論」, 丸善出版, 2022
[52]国立研究開発法人科学技術振興機構 研究開発戦略センター. 「人工知能と科学～AI・データ駆動科学による発見と理解～」, 戦略プロポーザル, CRDS-FY2021-SP-03, https://www.jst.go.jp/crds/report/CRDS-FY2021-SP-03.html, 2021
[53]日本学術会議.「オープンサイエンスの深化と推進に向けて」, 学術会議提言, https://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-24-t291-1.pdf, 2020
[54]Hanchen Wang et al. “Scientific Discovery in the Age of Artificial

Intelligence," Nature, Vol. 620, No. 7972, pp.47-60, 2023
[55]Nathaniel C Hudson et al. "Trillion Parameter AI Serving Infrastructure for Scientific Discovery: A Survey and Vision," Proceedings of the 10th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT '23), arXiv:2402.03480, 2024
[56]Mohamed Khalifa and Mona Albadawy. "Using Artificial Intelligence in Academic Writing and Research: An Essential Productivity Tool," Computer Methods and Programs in Biomedicine Update, 5, https://doi.org/10.1016/j.cmpbup.2024.100145, 2024
[57]Kayvan Kousha and Mike A Thelwall. "Artificial Intelligence to Support Publishing and Peer Review: A Summary and Review," Learned Publishing 37, https://doi.org/10.1002/leap.1570, 2023
[58]文部科学省.「令和6年版科学技術・イノベーション白書：AIがもたらす科学技術・イノベーションの変革」, https://www.mext.go.jp/b_menu/hakusho/html/hpaa202401/1421221_00020.html, 2024
[59]日本学術会議.「未来の学術振興構想（2023年版）：グランドビジョン⑩：データ基盤と利活用による学術界の再構築」, https://www.scj.go.jp/ja/info/kohyo/kohyo-25-t353-3.html, 2023
[60]Debleena Paul, Gaurav Sanap, Snehal Shenoy, Dnyaneshwar Kalyane, Kiran Kalia, Rakesh K Tekade. "Artificial Intelligence in Drug Discovery and Development," Drug Discovery Today, 2021 Jan;26(1), pp. 80-93, doi:10.1016/j.drudis.2020.10.010, 2021
[61]内閣府総合知ポータルサイト, https://www8.cao.go.jp/cstp/sogochi/index.html
[62]Tom Hope, Doug Downey, Daniel S. Weld, Oren Etzioni, Eric Horvitz. "A Computational Inflection for Scientific Discovery," Journal Communications of the ACM, Vol. 66, No. 8, pp.62-73, https://doi.org/10.1145/3576896, 2023
[63]Hiroaki Kitano. "Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery," AI Magazine, Vol. 37, No. 1, pp.39-49, https://doi.org/10.1609/aimag.v37i1.2642, 2016
[64]Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, David Ha. "The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery," http://arxiv.org/abs/2408.06292, 2024
[65]UNESCO. "Guidance for Generative AI in Education and Research," https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research#main-content, 2023
[66]文部科学省初等中等教育局.「初等中等教育段階における生成AIの利活用に関するガイドライン（Ver.2.0）」, https://www.mext.go.jp/content/20241226-mxt_shuukyo02-000030823_001.pdf, 2024

[67]Duri Long and Brian Magerko. “What Is AI Literacy? Competencies and Design Considerations,” Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp.1-16, 2020
[68]Hiroyuki Nitto, Daisuke Taniyama, Hitomi Inagaki. “Social Acceptance and Impact of Robots and Artificial Intelligence -Findings of Survey in Japan, the U.S. and Germany-,” NRI Papers, No. 211, https://www.nri.com/en/knowledge/report/lst/2017/cc/papers/0201, 2017
[69]細坪護挙・角田英之・加納圭・岡村麻子・星野利彦.「科学技術に関する国民意識調査－新技術の社会受容性－」, NISTEP RESEARCH MATERIAL, No.296, 文部科学省科学技術・学術政策研究所. https://doi.org/10.15108/rm296, 2020
[70]Makoto Nakada, Iordanis Kavathatzopoulos, Ryoko Asai. “Robots and AI Artifacts in Plural Perspective(s) of Japan and the West: The Cultural-Ethical Traditions Behind People’s Views on Robots and AI Artifacts in the Information Era,” The Review of Socionetwork Strategies, Vol.15, No.1, pp.143-168, 2021
[71]国立研究開発法人科学技術振興機構 研究開発戦略センター,「次世代 AI モデルの研究開発」, 戦略プロポーザル, CRDS-FY2023-SP-03, https://www.jst.go.jp/crds/report/CRDS-FY2023-SP-03.html, 2024
[72]The International Network of AI Safety Institutes. “Mission Statement,” https://aisi.go.jp/assets/pdf/Mission_Statement_International_Network_of_AI_Safety_Institutes.pdf, 2024
[73]内閣官房.「デジタル田園都市国家構想実現会議（第３回資料７）：デジタル人材の育成・確保に向けて」, https://www.cas.go.jp/jp/seisaku/digital_denen/dai3/siryou7.pdf, 2022
[74]岡本千草.「日本のスタートアップ事業分野とその立地パターンについて」, CREI Report, No. 16, https://www.crei.e.u-tokyo.ac.jp/wp-content/uploads/2024/03/537684ccd8af99d4bbc209d38e9f8e7d.pdf, 2024
[75]美馬のゆり.「生成 AI と教育：3. AI 時代を生きるリテラシーを育む－議論を基盤とした学習と問題解決型学習の新展開－」, 情報処理, Vol. 65, No. 7, e14-e19, 2024
[76]大学共同利用機関法人情報・システム研究機構 国立情報学研究所. 大学等におけるオンライン教育とデジタル変革に関するサイバーシンポジウム「教育機関 DX シンポ」開催一覧 アーカイブ, https://www.nii.ac.jp/event/other/decs/past.html

＜参考資料＞審議経過

令和６年

４月 22 日 情報学委員会（第 26 期・第８回）
提言の構成について

７月５日 情報学委員会（第 26 期・第９回）
提言骨子案について

10 月 22 日 第三部会（第 26 期・第４回）
委員会等からの活動報告（意思の表出検討中の委員会等）
情報学委員会（第 26 期・第 10 回）
提言案について

令和７年

２月 27 日 日本学術会議幹事会（第 381 回）承認