AI セーフティに関する具体的な影響の

調査報告書

令和 7 年 10 月 31 日

AI セーフティ・インスティテュート

調査委託先：NRI セキュアテクノロジーズ株式会社

AISI Japan AI Safety Institute

目次

1　調査の背景と目的

AI は急速な発展と普及を遂げており、AI が社会に与える影響もますます大きくなっている。そのような中、AI セーフティ・インスティテュート（以下「AISI」という。）では、2024 年 9 月に「AI セーフティに関する評価観点ガイド」（以下「評価観点ガイド」という。）及び「AI セーフティに関するレッドチーミング手法ガイド」（以下「レッドチーミング手法ガイド」という。）の 2 種類の AI セーフティに関するガイドを発行した。また、2025 年 3 月にはこれらのガイドの改定も実施した。これまで、これらのガイド作成にあたっては、AI システムがエンドユーザーに与える直接的な影響に焦点を当てた調査を実施してきた。しかし、AI システムがもたらす影響は個々のエンドユーザーにとどまらず、制度や社会システム、産業全体にも波及する社会技術的な領域に広がりつつある。ここでの「社会技術的」とは、AI 自体やそれを実装した AI システムに関する技術的要素と、AI や AI システムを取り巻く社会的要素の相互作用に着目する考え方である。近年、各国の AI に関連するガイドライン等において、AI や AI システムの社会技術的側面への言及がされることがある。例えば、米国 NIST の AI Risk Management Framework では、AI システムは本質的に社会技術的なものであり、AI のリスクとベネフィットは、システムの使用方法、他の AI システムとの相互作用、運用者、そしてシステムが導入される社会的背景に関連する社会的要因と組み合わされた技術的側面の相互作用から生まれる可能性があるとしている [1]。また、国際的な専門家の協力により作成された International AI Safety Report 2025 においても、単に技術的なアプローチにだけ焦点を当てるのではなく、社会技術的なシステムとして実装することが、AI システムによる危害を特定、調査、防御するために重要であることを指摘している [2]。

このような事情を踏まえ、本調査では AI の社会技術的な側面に着目する。つまり、AI やそれが実装された AI システムが社会における要素と相互作用し、実社会にどのような影響を及ぼしうるのかを調査する。また、調査を通して、AI の社会技術的な影響のうち、特に本調査時点において社会に大きな影響を与えている、あるいは与えようとしているものの実態を明らかにすることを目指す。また、調査で得られた内容を踏まえ、今後の AI セーフティに関連する施策に関連する示唆を見出すことを目指す。

---

[1] National Institute of Standards and Technology “Artificial Intelligence Risk Management Framework (AI RMF 1.0)” https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[2] Department for Science, Innovation and Technology and AI Safety Institute “International AI Safety Report 2025“ https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

2　調査の方法

本調査は、AI システムのうち、マルチモーダル情報を扱う基盤モデルを含む AI システムを主な対象とした。また、AI システムが及ぼす影響としては、AI システムがそのシステムのエンドユーザーに直接及ぼし得る影響に加え、AI システムのエンドユーザーを超えて、周囲の人々や社会に与える影響を対象とした。

本調査では、公開情報に基づく文献調査を実施し、その後、AI に関する社会技術的影響に関連し得るステークホルダーへのインタビュー調査を実施した。また、文献調査では、概要調査を実施した後に詳細調査を実施した。

各調査の方針は以下の各節において記載することとするが、各調査における目的は以下の通りである。まず、文献調査のうち概要調査では、AI セーフティに関する社会技術的影響の現状を概観できる事例や研究内容を収集し、詳細に調査するべき事項を把握することを目的として設定した。次に、文献調査のうち詳細調査では、概要調査の対象のうち特に重要と考えられる対象について、社会への影響や、影響するステークホルダー等を明確にすることを目的とした。また、収集した個別の事例について、本調査報告書で記載するのに適する粒度に一般化したトピックを導出することも目指した。インタビュー調査では、文献調査において導出したトピックのうち特に重要なものについて、国内における実態を詳細に把握することを目的とした。各調査における目的と各調査の関係について、図 1 に示す。

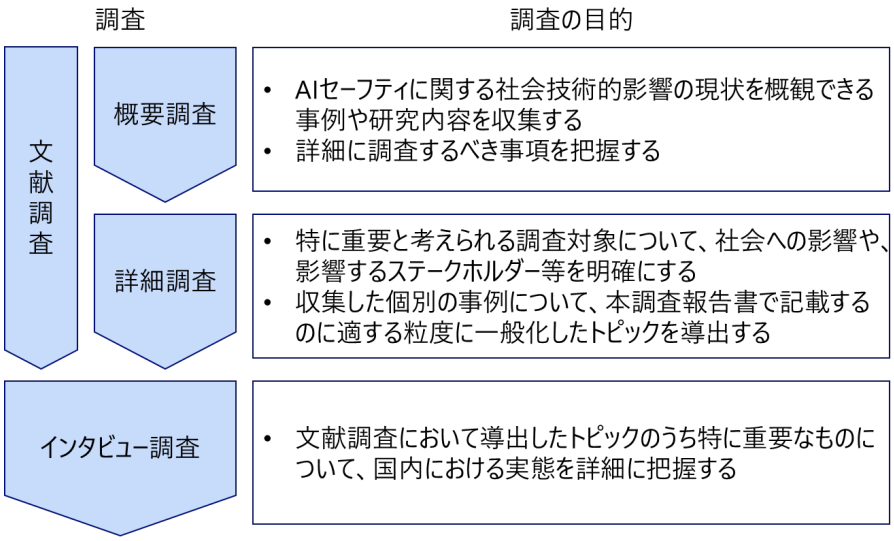

図 1 各調査の目的および関係性

2.1　文献調査

2.1.1 概要調査

文献調査では、まず、AI セーフティに関する社会技術的影響の現状を概観し、詳細に調査するべき対象を把握するため、概要調査を行った。国内外で発生した、あるいは今後発生する可能性がある、AI セーフティに関する社会技術的な影響について、国内外の文献（報道、学術論文、レポート）を対象にした概要調査を実施した。

概要調査では、AI セーフティに関連する社会技術的影響についての学術論文を参照することで導出したキーワード（偽情報、労働環境、環境等）を用いて公開情報を探索し、AI セーフティに関する社会技術的な影響に関連する文献（報道、学術論文、レポート）を収集した。収集は国内外をスコープに入れ、50 件程度の文献を収集した。なお、本調査時点において実際に社会に大きな影響を与えている事例を重点的に調査するために、7 割程度を報道の事例とした。収集した報道、学術論文、レポートについて、文献の趣旨、社会技術的な影響、文献に関するその他を整理した。

2.1.2 詳細調査

概要調査対象のうち、特に重要であると考えられる調査対象について、社会に与える影響や、影響を及ぼすステークホルダー等を明確にするため、詳細調査を行った。50 件程度の概要調査対象のうち、影響度・発生可能性・AI 固有と言える度合いが高いと考えられる 20 件文献を詳細調査対象として選定した。対象とした 20 件について、個別の事例に限らない項目として捉えられるように、適度に一般化し、項目の粒度を調整した。なお、詳細調査では、社会に与える影響、影響を及ぼすステークホルダー等について、概要調査において参照した文献以外にも、事例に関連する国内外の文献も参照して調査した。

さらに、対象とした 20 件について、内容の粒度や重複関係を考慮し、調査項目を 12 件の「調査トピック」として再構成した。3 章の調査結果は、この調査トピックの単位で記載した。

2.2 インタビュー調査

調査トピックのうち、特に重要であると考えられるものについて、国内における実態を詳細に把握するため、インタビュー調査を行った。詳細調査対象のうち、特に影響度・発生可能性・AI 固有と言える度合いが高く、直接的な影響を被る組織が存在すると考えられる、あるいは影響が特に国内において発生していると考えられる調査トピックを対象として 5 件選定した。選定した 5 件のトピックについて、インタビュー調査対象を 2 組織ずつ（計 10 対象）選定した。つまり 10 組織に対するサンプリング調査となり、本書で記載する調査結果はインタビュー組織毎の見解となることに留意いただきたい。インタビュー対

象組織（業種）と対象組織から公開が許諾されたインタビュー議事は A.1 に記載している。各調査における調査対象と各調査の関係を図 2 に示す。

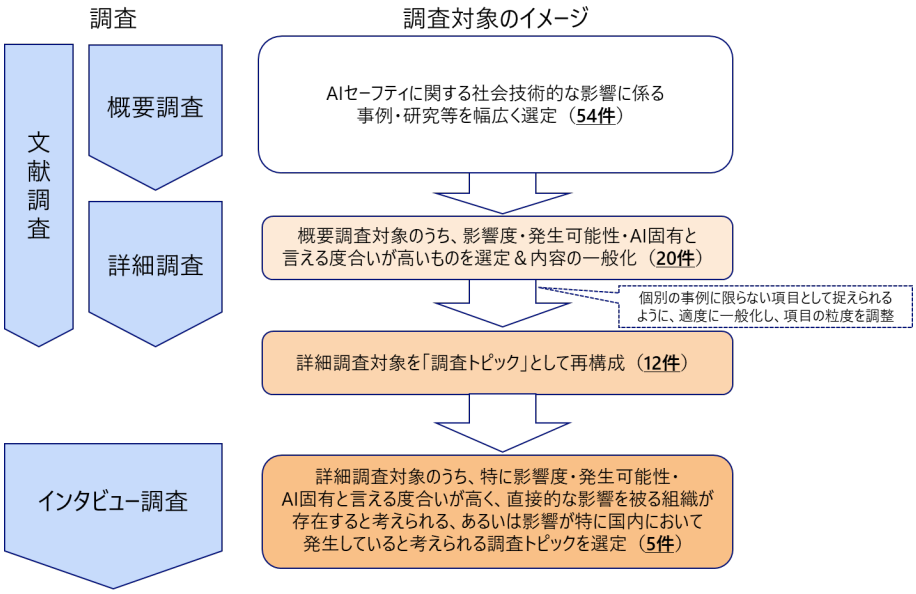

図 2 各調査の調査対象および関係性

3　調査結果

本章では、生成 AI の普及に伴う社会技術的な影響に関する調査結果を記載する。調査結果は、「倫理・法」、「経済活動」、「情報空間」、「環境」という 4 つの分類毎に整理して記載する。この分類は、経済産業省・総務省が発行する「AI 事業者ガイドライン（第 1.1 版） 別添」の「表 3. AI によるリスク例の体系的な分類案」のうち、「社会的リスク」に該当する中分類を援用したものである。AI 事業者ガイドラインにおいて規定されている社会的リスクと本事業において扱う社会技術的な影響は必ずしも一致する概念ではない。しかし、両者とも、AI という技術が社会に対して如何に影響を及ぼすかという点を扱っているという点では目的を共有するものである。そのため、本事業で調査を整理するに際して本分類を活用することとしている。表 1 に、「AI 事業者ガイドライン（第 1.1 版） 別添」における社会的リスクの中分類と、各分類に該当する調査トピックを示す。2.1.2 において記載した通り、調査トピックは文献調査における詳細調査の結果導出された項目である。本章の各節では、これらの調査トピックを単位として調査結果を記載している。

表 1 本報告書に調査結果を記載している調査トピック

分類	調査トピック
倫理・法への影響	生成 AI への過信が及ぼす心身への影響
	出力のバイアスが社会に及ぼす影響
	サイバー攻撃への悪用
	わいせつ物の生成や流通
経済活動への影響	生成物の権利に関する懸念
	労働環境への影響
	生成物の氾濫が及ぼす影響
	経済格差へ及ぼす影響
	機密情報の学習や漏洩
情報空間への影響	偽誤情報の生成や拡散
	多様性への影響
環境への影響	環境への影響

また、調査トピックについて、AI セーフティにおける重要要素との関連性を検討した。AI セーフティにおける重要要素とは、AISI が公開する評価観点ガイドにおいて示されている、AI セーフティを向上するうえで重視するべき重要要素である。具体的には、「人間中心」、「安全性」、「公平性」、「プライバシー保護」、「セキュリティ確保」、「透明性」の 6 つの要素が挙げられる。本調査におけるトピックは、AI セーフティにおける重要要素のいずれかに該当するものである。以下の表 2 は、調査トピックと AI セーフティにおける重要要素との関連性を示すマトリクス表である。表中で「●」が記載されている場合、該当する調査トピックと AI セーフティにおける重要要素が関連することを表している。

表 2 本調査における調査トピックと
AI セーフティにおける重要要素との関連性を示すマトリクス

		AI セーフティにおける重要要素					
		人間中心	安全性	公平性	プライバシー保護	セキュリティ確保	透明性
調査トピック	生成 AI への過信が及ぼす心身への影響	●	●				●
	出力のバイアスが社会に及ぼす影響	●		●			●
	サイバー攻撃への悪用		●			●	
	わいせつ物の生成や流通	●	●		●		●
	生成物の権利に関する懸念				●		●
	労働環境への影響	●					
	生成物の氾濫が及ぼす影響						●
	経済格差へ及ぼす影響	●		●			
	機密情報の学習や漏洩				●	●	
	偽誤情報の生成や拡散	●	●				●
	多様性への影響	●		●			
	環境への影響	●					

なお、AI セーフティに関する社会技術的な影響は、AI に関する技術的な環境の変化や、それを取り巻く社会的な環境の変化の影響を受けて急速に変化し得る。そのため、調査トピックの分類やそれらの内容については、継続的な検討を行うことが重要である。また、本調査では 2 章で示した通り、多くの分野における事例及び専門的知見を対象とした文献調査や、生成 AI に関する取り組みの実態を把握している組織へのインタビューを実施している。これにより、AI セーフティに関する社会技術的な影響に関して可能な限り多面的かつ実態に即した把握を試みている。もっとも、AI を取り巻く技術や社会情勢は急速に変化しており、それに伴い調査対象とすべきトピックや個々の調査項目も日々拡張している。そのため、本調査における調査トピックや各トピックにおける内容は、現時点において網羅性を完全に確保するものではないことに留意が必要である。

以下の各節では、AI セーフティに関する社会技術的な影響に係る調査トピックを「倫理・法」、「経済活動」、「情報空間」、「環境」の分類に沿って記載する。また、各調査トピックについて、トピックの概要説明、文献調査結果、インタビュー調査結果を記載している。次節以降における、調査トピック毎の記載項目及び記載内容を表 3 に示す。

表 3 各調査結果の記載項目

項目	記載内容
トピックの概要説明	該当する AI セーフティに関する社会技術的な影響がどのようなものなのか、また、なぜ重要と考えられるのかを記載する。
文献調査結果	各トピックに関する調査実施時点での状況について、公開情報に基づいた調査結果を記載する。具体的には、本トピックに関連して、社会への影響が現状どの程度、どのように生じると想定されているか、どのようなステークホルダーが本トピックに関係しているか、について記載する。また、A.1 において、概要調査対象の一覧を記載する。なお、報道は「倫理・法」、「経済活動」、「情報空間」、「環境」で分類しているが、学術論文、レポートは 1 文献内に複数の事例について言及されているため、分類していない。
インタビュー調査結果	各トピックのうち、より詳細に実態を把握することが重要であると考えられたものについて、インタビュー調査を実施した結果を記載する。具体的には、インタビュー対象組織において認識されている、各トピックに関する現状の影響や今後の影響に関する理解、各影響への対応に関する考え方について記載する。各節においてインタビュー対象組織の概要を記載する。また、A.2 において、本事業で実施したインタビュー調査の対象組織（業種）一覧とインタビュー対象から公開が許諾されたインタビュー議事を記載する。

3.1 倫理・法への影響

3.1.1 生成 AI への過信が及ぼす心身への影響

■ トピックの概要説明

生成 AI が技術的に発展し、生成物の精度が飛躍的に向上したことで、日常的な情報検索、メール文案作成、アイデア出しの壁打ち等、これまで人間が役割を担っていた用途に活用されるようになり、社会に広く普及している。ただし、そのような利便性の一方で、生成 AI を過信することで、心身に影響がなされないかという懸念も一部で指摘されている。

ここでは、生成 AI への過信がもたらす社会技術的な影響として、不適切な誘導が心身に及ぼす影響、および思考力に及ぼす影響という 2 つの影響を取り上げる。生成 AI の利用はタスクの効率化や高品質化といった価値だけでなく、話し相手や心の支えといった情緒的価値を求めるエンドユーザーにも広がっている。しかし、生成 AI に依存するエンドユーザーに対して不適切な誘導が行われた場合、最悪の場合自殺に至ってしまうケースも確認されている。また、学生が生成 AI を過信することで、思考力に影響を及ぼす可能性があることが、国内外の研究によって指摘されている。これらの問題は単なる技術的な問題ではなく、社会全体に波及する重要な社会技術的問題と考えるため、本事業における検討対象としている。

■ 文献調査結果

生成 AI への過信が及ぼす心身への影響に関する調査からは、以下のような知見が得られている。

まず不適切な誘導に関して、複数の事例が報告されている。2021 年に英国で発生した事件では、男が AI チャットボットと 5,000 件以上のメッセージを交わし、犯行を後押しされる形でエリザベス女王の暗殺未遂を起こした [3]。また、2023 年にはベルギー人男性が AI チャットボット「Eliza」と 6 週間の対話を続けた後に自殺しており、チャットログには「天国で一つになる」といった生成 AI による肯定的誘導の痕跡が残されていた [4]。さらに 2024 年には米国フロリダ州で 14 歳の少年がチャットボットとの長期的対話に依存した結

[3] 読売新聞,「「殺し屋でも愛してくれるか」「もちろんです」…ＡＩの恋人、女王殺害を後押し」 https://www.yomiuri.co.jp/world/20240212-OYT1T50014/

[4] Japan Broadcasting Corporation (NHK),「生成 AI と会話を続けた夫は帰らぬ人に…」 https://www3.nhk.or.jp/news/html/20230728/k10014145661000.html

果、自殺に至った事例も報告されている[5,6,7]。これらのケースは、生成 AI がエンドユーザーの感情的依存を強化し、不適切な誘導を行った際に現実の行動へと直結するリスクを示している。

こうした事例は偶発的なものではなく、エンドユーザーの心理状態や AI 設計の特性に起因する構造的なリスクであると考えられる。株式会社電通が 2025 年に実施した調査[8]では、全国 1,000 人のうち約 65%が「AI に感情を共有できる」と回答し、10 代の 41.9%が週 1 回以上 AI と対話していると報告されている。特に若年層において、AI を「心の支え」や「話し相手」と位置づける傾向が強く、依存リスクが高いことが確認された。欧州最大の科学技術分野における研究機関である Fraunhofer Institute for Production Systems and Design Technology (Fraunhofer IPK)に所属する研究員の Vivek Chavan ら[9]は、感情表現豊かな生成 AI は伴侶的役割を果たす一方、過度な依存を生みやすいと警告しており、エンドユーザーへ反論しない設計や常時利用可能である仕様が依存形成を助長すると指摘している。

次に、生成 AI が思考力に及ぼす影響についても多くの研究が報告されている。杭州師範大学によるメタ分析では、生成 AI を「インテリジェント・チューター」として利用する場合に高次思考力が向上するとの結果が得られている。これは個別化されたフィードバックを通じて学習者の内省を促進する効果によるとされる[10]。他方で、Massachusetts Institute of Technology (MIT) Media Lab の研究では、生成 AI に依存した場合に脳活動や独自性が低下し、批判的思考力が削がれる「認知的負債（Cognitive Debt)」が蓄積されることが示されている[11]。さらに、ブレーメン大学の調査では、生成 AI を使用してレポートを作成した学生は非使用者よりも期末試験の点数が平均 6.7 点低下しており、特に優秀

[5] New York Times, “Can a Chatbot Named Daenerys Targaryen Be Blamed for a Teen’s Suicide? “
https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html

[6] Cable News Network (CNN), “There are no guardrails.’ This mom believes an AI chatbot is responsible for her son’s suicide“
https://edition.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit

[7] Yahoo!ニュース, 「「AI とのチャットに依存、14 歳が死亡」母親が提供元を提訴、その課題とは？」
https://news.yahoo.co.jp/expert/articles/7225ddf3ec2e66fae6a09bd6cc96313b2a44e6f8

[8] 株式会社電通, 「「対話型 AI」に感情を共有できる人は 64.9% 「親友」「母」に並ぶ"第 3 の仲間"に」
https://www.dentsu.co.jp/news/release/2025/0703-010908.html

[9] Chavan, Vivek, et al. "Feeling Machines: Ethics, Culture, and the Rise of Emotional AI." arXiv preprint arXiv:2506.12437 (2025).
https://arxiv.org/pdf/2506.12437

[10] Nature HSS Communications, “Meta-analysis of ChatGPT impact on learning outcomes and higher-order thinking“
https://www.nature.com/articles/s41599-025-04787-y

[11] Massachusetts Institute of Technology (MIT) Media Lab, “Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task“
https://www.media.mit.edu/publications/your-brain-on-chatgpt/

な学生ほど悪影響が顕著であるとされた[12]。

このように、生成 AI は利用方法次第で学習促進効果も期待できる一方、過度の依存や無批判な使用は思考力低下を招くリスクを孕んでいる。Microsoft Research と Carnegie Mellon University (CMU)の共同研究では、生成 AI を信頼する度合いが高い人ほど批判的思考が弱まることが確認され、AI 過信が知的活動の質を損なう可能性が実証的に示されている[13]。

次に、生成 AI の過信が心身に及ぼす影響に関連しうるステークホルダーについて考察する。本影響に関連しうるステークホルダーとして、例えば、AI 開発者、AI 提供者、エンドユーザー、教育機関・雇用者、政府機関、行政機関・国際機関が挙げられる。なお、AI 開発者、AI 提供者、AI 利用者の定義は AI 事業者ガイドライン（第 1.1 版） を参照いただきたい。AI 開発者は、生成 AI が心理的依存を強めたり、不適切な誘導を行ったりすることを防ぐために、開発段階から心理学や教育学の専門家を参画させ、依存兆候を検知できる機能を設計に組み込む必要があると考えられる。AI 提供者は、チャットボットサービスの運用において、不適切な誘導を防ぐ防御機構を備える責任を負うと考えられる。

教育機関や雇用者においては、学習者や知的労働者が主要な影響を受ける主体であるため、AI リテラシー教育や評価設計の整備が必要であると考えられる。政府機関は、制度面での対応が求められる。文部科学省はプロセス（学習過程）評価や口頭発表を組み込む教育方針を提示[14]しており、AI 依存を抑制し批判的思考を維持する教育制度の構築が課題としている。最後に行政機関や国際機関は、規制やガイドラインを通じ、心理的リスクを軽減しつつ健全な AI 活用を促進する役割を担う必要があると考えられる。

■　インタビュー調査結果

本トピックに関するインタビューは、教育事業者の社内シンクタンク A 社、AI 利用・統制の進む B 大学に対して実施した。A 社、B 大学ともに学生や教員の生成 AI 活用に関する実体験に基づくデータを有しており、外部発信等を行っているため、インタビューの対象組織として選定した。ここでは、インタビューから得られた要旨（回答者の発言）を記載することとし、各インタビューの詳細については A.2 のインタビュー議事を参照いただきたい。

●A 社に対するインタビュー調査において把握した事項

[12] Wecks, Janik Ole, et al. "Generative AI Usage and Exam Performance." arXiv preprint arXiv:2404.19699 (2024). https://arxiv.org/pdf/2404.19699

[13] Microsoft Research, “The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers“ https://www.microsoft.com/en-us/research/wp-content/uploads/2025/01/lee_2025_ai_critical_thinking_survey.pdf

[14] 文部科学省,「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」 https://www.mext.go.jp/content/20241226-mxt_shuukyo02-000030823_001.pdf

A 社は通信・教育事業を担う事業会社内の社内シンクタンクであり、学生・教員における生成 AI の活用・影響などについて、調査・分析を実施し、その知見を自社メディアや報道機関を通して広く公開することで、教育リテラシーの向上等に広く貢献している。A 社の見立てでは、生成 AI の教育現場への浸透度は学齢が上がるほど高い。大学生、とりわけ就活期の上級生はレポート作成や企業へのエントリーシート下書きに日常的に AI を活用しており、専門学校や芸術系大学でも利用事例がある。高校生は検索手段として ChatGPT を用いるようになっている。また、先生の判断で情報の授業で AI を利用し、動画や文書作成等様々な取り組みがなされているという。中学校は保護者許可の下で翻訳や調べ学習に対して試行的に導入している事例があり、小学校は授業より教員の校務文書作成での利用が中心で、児童向け学習ユースケースはまだ少ないとされている。

生徒が使う AI の機能は情報整理、文章構成、画像・動画生成など初歩的なものが大半だが、将来は生徒の学習ログを可視化し個別最適化指導へつなげる「自学＋教員の伴走」型の活用が期待される。また、AI を利用することで生徒の長文作成の苦手意識が減る一方で「自分で考える」機会が減る懸念もなされているが、AI を利用するメリットが大きければ問題ないとの見方もあるという。思考力への影響としては、ワンクリックでクリエイティブな作業を行うことができるようになった結果、自分で試行錯誤した経験のある者は見極め力を発揮できるが、その経験を欠く者は判断力を養えないリスクがある。また、その他のテクノロジーと同様、意欲と継続力を持つ生徒が先へ進み、そうでない層が取り残され格差が拡大する可能性がある。

今後 AI を利用した教育の効果を最大化する鍵は「管理から自律へ」の転換だと A 社は指摘する。一人ひとりが AI をカスタマイズし、自分の動機に基づいて学ぶ環境づくりが重要で、その前提として日本にある「皆と同じが正しい」という風潮から脱し、科学的・論理的に主体的判断を下す姿勢を育てる必要がある。

●B 大学に対するインタビュー調査において把握した事項

B 大学では、生成 AI 利用に関する方針を明確化の上、大手生成 AI サービス提供ベンダと法人契約を結んで有料版サービスを全学生に提供している。大学における生成 AI 利用のルールとして、個人情報や研究データは大学提供の生成 AI 環境でのみ利用するルールとなっている。大学内における具体的な利用方法としては、英語授業の翻訳やルーター設定のエラー対応など、ティーチングアシスタント的に活用されている。また、卒論・レポート作成時の文章のブラッシュアップや、モックアップ・デジタルプロトタイプの作成、資格学習の活用等、生成 AI を用いた生成物の品質向上や短期間での作成といった形での利用が増えてきている。

生成 AI が学生の思考様式に与えるプラス面として、AI に質問しながら適切に言語化する過程で「正しく聞く力」が鍛えられ、ファクトチェックを通じて本質的な議論が可能になるとの言及があった。AI の出力を無批判に受け入れるだけでは能力は伸びないが、結果

を吟味し再構築する学生は思考力が向上すると期待できる。したがって、AI 活用は思考深化の契機となり得る。一方でネガティブな側面として、学生が AI の出力内容を理解せずに利用するケースがあり、特に研究室の進捗報告書などで誤情報が混入する危険が指摘された。本ケースへの対策としては、教員が質問を投げかけて学生の理解度を確認し、AI が提供した情報を自ら検証・説明させる指導が有効とされる。批判的・論理的思考力の低下については、現時点で明確なエビデンスは乏しいものの、短時間で異常に高品質なレポートが提出される事例が見られ、AI 支援が過度になると思考プロセスが省略される恐れが示唆された。これらは個別事例であり、体系的な調査が必要だが、教師側が AI 利用の前提を明示し、結果の裏付けや根拠説明を求めることで、思考力低下リスクを抑制できると結論付けられた。以上のように、生成 AI は思考の深化と同時に誤用リスクを孕むツールであり、適切なリテラシー教育と教員の質問・フィードバックを組み合わせた指導が、ポジティブな効果を最大化しつつネガティブな影響を最小化する鍵となる。

教育分野での AI 活用として、ハンズオントレーニングやプログラム開発のアシスタントとしての活用が、人間を上回る効率を発揮でき効果的であると言及されている。また、B 大学内には AI センターが設置されており、抽象的・高度なテーマの教育活用を議論する動きも進んでいる。特に博士課程レベルの知識を備えた生成 AI は、論文執筆支援に有効で、誤字脱字の自動修正や一次スクリーニングを行うことで、教員と学生が本質的な議論や批判的検討に費やす時間を増やすことができる。他方で、教育分野への適用における課題としては、利用できる有料サービスの有無による学習格差が挙げられる。B 大学は契約で格差を解消したが、未契約校や初等・中等教育で AI に触れた生徒と触れない生徒の間でも差が拡大する恐れがある。現在の教育は、基礎的な計算力や漢字習得に偏重した AI 活用を想定していない教育となっていることも課題視されている。これら課題の解決として、政策面で AI を教育現場に導入できる環境整備を行い、無償教育の一環として PC 配布と同じように AI 利用環境の提供を盛り込むべきだと提言された。これにより、全学年・全学部で均等に AI リテラシーを養い、格差を最小化しつつ、生成 AI の利点を最大限に活かした次世代教育が実現できるのではないかとの言及があった。

●A 社、B 大学のインタビュー調査結果を踏まえた考察

上記の A 社、B 大学へのインタビューの結果から、既に教育の現場で AI が活用されつつあり、特に大学機関において活用が進んでいることが確認できる。思考力への影響については、学生が AI の出力内容を理解せずに利用することによる批判的・論理的思考力の低下が懸念されるが、適切なリテラシー教育や AI の利用を前提とした教育プログラムの設計等を行うことでポジティブな効果を最大化しつつネガティブな影響を低減させることができると考えられる。

3.1.2 出力のバイアスが社会に及ぼす影響

■　トピックの概要説明

生成 AI は教育、医療、行政等幅広い分野での利活用が進んでいる。一方で、生成 AI の出力には学習データ由来のバイアス（偏り）が反映されることがある。このバイアスは、歴史的・社会的に形成された差別や不均衡を再生産する危険をはらみ、人種、性別、宗教、障害、文化的背景などに基づく不当な扱いやステレオタイプを助長する可能性が指摘されている。なお、このような出力が社会での多様性に及ぼす影響は、本報告書の多様性への影響（3.3.2 節）において扱う。ここでは特に、このようなバイアスが社会的な差別につながりうることについて扱うこととする。

生成 AI の出力におけるバイアスの問題は、単なる技術的な問題に留まらない。社会構造や制度に深く作用し、情報流通の公平性や少数者の包摂性といった社会的価値に影響を及ぼす恐れがあり、社会技術的影響として重要であると考えられる。特に採用や教育、行政判断といった人々の生活や権利に直結する領域において、生成 AI の出力にバイアスが含まれる場合、その影響は制度化・構造化され、長期にわたり社会的不利益を固定化する恐れがあると考えられる。以上より、出力のバイアスが社会に及ぼす影響は重要な社会技術的影響であると考え、本事業における検討対象としている。

■　文献調査結果

ここでの差別とは、性別、人種・民族、宗教といった社会的要素に基づき、そのような要素を持つ、あるいは持たない人々に対して不当と考えられる取り扱いをすることを指すこととする。以下では、性別、人種・民族、宗教の 3 つの社会的要素に関する差別が生成 AI の出力によるバイアスによって生じた事例や、関連する研究について記載する。

第 1 に、性別に関する差別について、国際連合教育科学文化機関 (UNESCO)が 2024 年に発表した調査では、主要な LLM である GPT-2、ChatGPT、LLaMA1 において、女性が「家庭」「子ども」といった語と強く結びつけられ、男性が「ビジネス」「キャリア」と関連づけられる傾向が確認された。さらに職業の文脈では、男性には「教師」や「医者」といった専門職が関連づけられる一方、女性には「家事使用人」や「料理人」といった役割が割り当てられる傾向が見られた [15]。このようなジェンダーバイアスは単なる表象の問題にとどまらず、推薦文の生成においても再生産されており、スタンフォード大学が開発した Alpaca では、男性に対して「専門性」「誠実さ」といった表現が用いられるのに対し、女性に対しては「美しさ」「喜ばせる」といった言葉が付与されるケースが報告され

[15] 国際連合教育科学文化機関 (UNESCO), “Challenging systematic prejudices: an investigation into bias against women and girls in large language models“
https://unesdoc.unesco.org/ark:/48223/pf0000388971

ている[16]。

第2に、人種・民族に関する差別についても事例が確認されている。認知科学者であるBirhaneらは、マルチモーダルデータセットの規模拡大に伴い、黒人やラテン系の人々が犯罪者として誤って分類される可能性が高まることを指摘した[17]。またスタンフォード大学の研究では、アフリカ系アメリカ英語の話者が標準アメリカ英語の話者と比較して低い地位の職業に割り当てられたり、架空の刑事裁判のシナリオにおいてより重い罰を適用したりする傾向があるとしている[18]。こうした結果は、使用する言語や方言そのものが差別の根拠とされるリスクを示している。

第3に、宗教に関する差別について、主に実証的な研究における指摘が確認されている。スタンフォード大学在籍でAI・機械学習の研究者であるAbidら（2021）の調査では、GPT-3が「イスラム教徒」を23%のケースで「テロリスト」と結びつけ、「ユダヤ人」を5%のケースで「お金」と関連づけて描写していたことが確認された[19]。このようなバイアスは宗教に対する固定観念を助長し、社会的な排除や偏見を強化する危険性を持っている。

加えて、複合的な差別の事例も報告されている。画像生成アプリLensaでは、アジア系女性の画像が過度に性的な描写として生成されやすい傾向があり、表象的被害が顕著に現れている[20]。また、ワシントン大学の研究によれば、履歴書評価において白人男性の名前が最も高い確率で採用に適すると判断され、女性や黒人男性の名前は好ましいとされる比率が著しく低かったことが判明した[21]。この結果は、生成AIの出力が採用や人材評価といった社会的に重要な意思決定に組み込まれた場合、差別を制度化・構造化してしまう可能性を示唆している。

次に、生成AIによる出力のバイアスが社会に及ぼす影響に関連しうるステークホルダ

---

[16] Wan, Yixin, et al. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters." arXiv preprint arXiv:2310.09219 (2023).
https://arxiv.org/abs/2310.09219

[17] Birhane, Abeba, et al. "The dark side of dataset scaling: Evaluating racial classification in multimodal models." Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024.
https://arxiv.org/pdf/2405.04623odels

[18] Stanford HAI, “Covert Racism in AI: How Language Models Are Reinforcing Outdated Stereotypes”
https://hai.stanford.edu/news/covert-racism-ai-how-language-models-are-reinforcing-outdated-stereotypes

[19] Abid, Abubakar, Maheen Farooqi, and James Zou. "Persistent anti-muslim bias in large language models." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021.
https://arxiv.org/pdf/2101.05783

[20] MIT Technology Review, “The viral AI avatar app Lensa undressed me—without my consent“
https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/

[21] University of Washington, “AI tools show biases in ranking job applicants’ names according to perceived race and gender“
https://www.washington.edu/news/2024/10/31/ai-bias-resume-screening-race-gender/

ーについて考察する。本影響に関連しうるステークホルダーとして、例えば、AI 開発者、AI 提供者、AI 利用者が挙げられる。 AI 開発者は、バイアスの発生源となり得る学習データの収集・前処理・モデル設計を技術的に制御できる立場にある。さらに、生成 AI が採用判断や司法判断などに利用された場合には社会的に重大な影響が及び得ることから、AI 開発者には、結果として生じ得る被害の大きさに見合った一層の慎重性が求められる。AI 提供者は、生成 AI を用いたサービスにおける出力制御や利用規約の設定等を行う立場にある。そのため、バイアスによって生じる差別について強い関連性があるステークホルダーと言える。特に人事評価や画像生成といった分野では社会的影響が大きいため、責任が求められると考えられる。

さらに AI 利用者も大きな役割を果たすが、どのように生成 AI を使用するかによって関与の形は多様である。政府がバイアスを含む出力を意思決定に用いた場合、その偏りが制度として長期的に固定化される危険性がある。また、クリエイティブ職が偏った生成物をコンテンツ制作に組み込むことで、差別的な表現が無自覚のうちに広く流通し、文化的規範として再生産されてしまう可能性がある。そして、バイアスの影響を最も直接的に受ける少数者当事者もまた不可避的にステークホルダーであり、彼らの視点を欠いた AI システムの設計や運用は正当性を欠くことになる可能性がある。

3.1.3 サイバー攻撃への悪用

■　トピックの概要説明

生成 AI は本来、教育・医療・産業など幅広い分野で革新的な活用が進められているが、その汎用性ゆえにサイバー攻撃に悪用されるリスクが急速に高まっている。自然言語の生成、マルチモーダル処理、コード生成といった機能は、攻撃者にとって効率的な攻撃ツールとなり得る。これにより、従来は専門的な知識や労力を要したサイバー攻撃が自動化・効率化され、誰でも容易に実行できる状況へと変化している。

さらに、生成 AI を利用したサイバー攻撃は単なる技術的脅威にとどまらず、市民や企業の信頼基盤を揺るがし、制度や経済にも波及する社会技術的課題としての性質を有している。例えば、フィッシングやランサムウェア攻撃に加え、ディープフェイクを用いたなりすましや本人確認突破といった新しいサイバー攻撃の形態が登場しており、金融・行政・医療といった重要分野に直接的な影響を与えている。

上述のように、生成 AI のサイバー攻撃への悪用は、従来の情報セキュリティの枠を超え、技術・社会・制度が相互作用する領域において対策を検討することが必要なテーマであるため、本事業における検討対象としている。

■　文献調査結果

生成 AI のサイバー攻撃への悪用は、市民や企業をはじめとした社会全体に深刻な影響を与えている。

まず市民に対する影響として、従来のフィッシング詐欺では文法や言葉遣いの不自然さがサイバー攻撃の兆候となっていたが、生成 AI の発展によって極めて自然で流暢な言語表現が容易に生成できるようになり、この検出手法の有効性は急速に低下している。Proofpoint 社の調査によれば、2025 年 2 月時点で全世界の新種メール攻撃の 80%以上が日本を標的にしており、その多くで検知回避機能を備えたフィッシングキット「CoGUI」が使われている。さらに生成 AI による自然な日本語生成により詐欺メールを見抜きにくくなり、認証情報窃取やアカウント乗っ取りの被害につながっている [22]。

次に企業への影響としては、生成 AI によるサイバー攻撃のスケールと精度の飛躍的向上が指摘されている。CrowdStrike 社は生成 AI を用いたサイバー攻撃の特徴として、「攻撃の自動化」「効率的な情報収集」「攻撃内容の高度なカスタマイズ」「強化学習による進化」「従業員の個別標的化」の五点を挙げている [23]。こうした特徴を背景に、経営層を装った標的型攻撃や組織の権限者へのなりすましが容易となり、実際に香港では 2024 年、CFO に偽装したディープフェイク映像を用いたビデオ通話詐欺によって約 2,500 万ドルの

[22] 日本プルーフポイント株式会社,「日本が今、最も狙われている—急増する DDoS 攻撃とメール攻撃の実態」 https://www.proofpoint.com/jp/blog/email-and-cloud-threats/Japan-is-now-the-most-targeted-country-in-the-world

[23] CrowdStrike, “AI-Powered Cyberattacks” https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/ai-powered-cyberattacks/

不正送金が発生した事例が報告されている[24]。また、2024 年 5 月、日本国内で生成 AI を利用し、得られた設計情報を組み合わせてランサムウェアを作成した 25 歳の男が逮捕される事件が発生した[25]。さらに、金融分野においても生成 AI は深刻な影響を与えている。特に「合成身元詐欺（Synthetic Identity Fraud）」が問題視されており、これは実在する個人情報の断片と生成 AI によって作成された虚偽情報を組み合わせ、あたかも実在する人物であるかのように偽装して金融口座やクレジットカードを取得する手法である。Wakefield Research 社の調査によれば、対象企業の 87%が合成身元の顧客（実在の人物を装って作られた偽物の顧客）にクレジットを提供した経験を持ち、23%は 1 件あたり 10 万ドル以上の損失を被ったと回答している[26]。このような手法は金融システムの根幹を揺るがし、社会的信頼基盤を侵食するリスクを含んでいる。

次に、生成 AI によるサイバー攻撃への悪用で影響に関連しうるステークホルダーについて考察する。本影響に関連しうるステークホルダーとして、例えば、AI 開発者、AI 提供者、エンドユーザー、一般企業、政府・国際機関が挙げられる。AI 開発者はモデル設計や学習データ選定を通じて出力特性を左右する立場にあり、意図しない悪用への責任回避が難しくなりつつある。AI 提供者は出力制御や利用規約策定、コンテンツモデレーションを担う立場として、悪用防止に直接的な責任を負うと考えられる。エンドユーザーは被害者となる一方で、悪意あるエンドユーザーは攻撃者としてリスクを増幅させる存在でもある。また、一般企業は標的型攻撃や認証基盤の突破といったリスクにさらされている。加えて、政府や国際機関といったガバナンス主体も重要なステークホルダーであり、AI Safety Institute (AISI)としては AI セーフティに関する評価基準の策定や国際的な協調を進めることで、生成 AI が悪用されるリスクの防止に取り組んでいる。

■　インタビュー調査結果

本トピックに関するインタビューは、情報セキュリティベンダーC 社、D 社に対して実施した。C 社、D 社ともに外部の悪意あるエンドユーザーによるサイバー攻撃を想定した情報セキュリティソリューションを提供しており、サイバー攻撃への生成 AI の悪用に関する情報発信を積極的に実施していることから、インタビューの対象組織として選定した。ここでは、インタビューから得られた要旨（回答者の発言）を記載することとし、各インタビューの詳細については A.2 のインタビュー議事を参照いただきたい。

---

[24] Cable News Network (CNN), “Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’”
https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

[25] 読売新聞,「生成ＡＩ悪用しウイルス作成、警視庁が２５歳の男を容疑で逮捕…設計情報を回答させたか」
https://www.yomiuri.co.jp/news/national/20240528-OYT1T50015/

[26] Deduce, “Protection Against Synthetic Identity Fraud is Failing”
https://www.deduce.com/resource/wakefield-research-report/

●C 社に対するインタビュー調査において把握した事項

C 社は迷惑メール・マルウェア・ネット詐欺等のインターネット上の脅威に対して、法人・個人の双方に脅威を防ぐためのソリューションを提供している。また、その知見を自社メディアや報道機関を通して広く公開することで、情報セキュリティリテラシーの向上等に広く貢献している。

C 社が観測している脅威の実態として、生成 AI を悪用することによってサイバー攻撃が劇的に高度化したといった事象は確認されていない。他方で、攻撃の「量産化」や「効率化」の観点で、活用されていることが確認されている。特に生成 AI サービスの普及以後、フィッシング攻撃において、文言を少しずつ変えながら、より自然な言い回しの攻撃が大量に出回るようになっており、メールフィルタリング対策を突破しつつ、攻撃成功率を高める活動が実施されており、言語の壁によりフィッシング攻撃の難易度が高かった日本国等を中心に、攻撃の増加が観測されている。なお、本攻撃においては、あくまで自然な日本語の文面を作成しているだけのため、WormGPT のような犯罪専用ツールだけでなく、ChatGPT 等の正規の汎用サービスも悪用していると推察されている。また、新種の脅威として、LLM を利用して、感染端末上でリアルタイムに攻撃コマンドを生成するマルウェアが出現している。外部の公開 AI サービスに API 接続するもの、ローカル PC 内で起動するものの二種類が既に 2025 年夏時点で確認されている状況にある。ただし、現段階において、これらマルウェアはあくまで人間が実行するようなコードをリアルタイムで生成しているだけであり、従来と異なる高度な攻撃を実施しているわけではなく、まだ概念実証（PoC）の段階にある可能性が高いと推察されている。

AI を用いたサイバー攻撃への対策として、「巧妙化への対応」と「量とスピードへの対応」の必要性を述べられている。AI を悪意あるエンドユーザーが活用することで、人間が見抜くことが困難な攻撃が増えるため、AI を用いて映像の微細な変化や音声ノイズを分析・検知する技術が重要になると推察されている。また、攻撃が大規模化する一方、文面の変化などによりシグネチャベースの防御は有効ではなくなるため、防御側も AI を活用し、平時と異なる振る舞いをリアルタイムに検知・遮断する仕組みの重要性が増すと推察している。また、一般人は「攻撃が見破れないこと」を前提とした行動がますます重要になることを指摘しており、公式アプリやブックマークからアクセスする習慣を徹底する等のふるまいがより重要になると言及している。加えて産学官連携として、IP アドレス等の従来の脅威情報に加え、「悪用されたプロンプト」や「AI モデルの種類」といった AI の悪用を前提とした新しい情報を共有する枠組みの必要性とコンテンツの信頼性を証明する技術・枠組みの開発、および AI を悪用する行為に対する法整備の検討が必要であることに言及している。

●D 社に対するインタビュー調査において把握した事項

D 社は、迷惑メール・マルウェア等のインターネット上の脅威に対して、法人・個人の

双方に脅威を防ぐためのソリューションを提供している。また、その知見を自社メディアや報道機関を通して広く公開することで、情報セキュリティリテラシーの向上等に広く貢献している。

D 社が観測している脅威の実態として、生成 AI サービスの普及以後、フィッシングメールの急増が確認されている。具体的には、2024 年度のフィッシングメール件数は、2023 年度の件数の 7〜8 倍に達していることを観測している。また、言語の壁で守られていた（言語の違いにより海外からの攻撃が成立しづらかった）日本を代表とする国々が主な標的となっていることを観測しており、2024 年度の新規フィッシングメールのうち、約 8 割が日本語で記載されたものとなっている。なお、フィッシングメールにはエンドユーザーを標的とするもの、企業の従業員を標的とするもの、双方を標的とするもの（判断できないもの）が存在するが、いずれも同程度の増加を観測しており、生成 AI の普及に伴い、特定の標的に対する攻撃のみが増加したとの事実は観測されていない。また、フィッシングメールによって取得する情報も認証情報の窃盗が主流であり、従来のフィッシングメールと攻撃の狙いが大きく変化はしていない。また、生成 AI を活用した新種の攻撃として、社長等の経営層の声を生成 AI に学習させ、不正な送金を指示するボイスフィッシング攻撃が確認されている。

生成 AI を用いたサイバー攻撃への対策として、攻撃の「量とスピードへの対応」の必要性が述べられている。具体的には、なりすましメール対策（DMARC 等）や電話番号の詐称対策（STIR/SHAKEN 等）等のグローバルスタンダードの技術対策を導入した上で、強固な本人確認を実現する技術（eKYC 等）や生成 AI を使った不正検知等の対策を実施することが推奨されている。加えて、より防御力を高めるために、攻撃手法に関する迅速な情報共有が重要であり、システム間で連携できる仕組みの必要であると述べている。また、AI エージェントの普及に伴い、データへのアクセス権限の管理がより重要となるため、フィッシング攻撃などを介して過剰な権限を与えてしまうことないよう、AI を使いこなす人材へのセキュリティ教育が新たな課題となると述べている。

●C 社、D 社のインタビュー調査結果を踏まえた考察

上記の C 社、D 社へのインタビューの結果から、既にサイバー攻撃に生成 AI が悪用されており、攻撃量の増加と攻撃品質の向上が達成されていることが確認できる。また、言語の壁で守られていた日本が、喫緊最も被害を最も受けている国の一つとなっていることが確認できる。これら攻撃者側の環境変化を踏まえ、企業・エンドユーザー側は対策をより強化していく必要があり、基本的な対策の実施に加えて、脅威情報共有等の既存の枠組みのアップデート、および生成 AI を活用したソリューションの導入を検討していく必要があると考えられる。

3.1.4 わいせつ物の生成や流通

■　トピックの概要説明

生成 AI の急速な発展により、従来は人手による編集や高度な技術を要したわいせつ物の生成が、誰でも容易に行えるようになっている。その結果、児童や女性の権利侵害、性的搾取といった深刻な社会課題が顕在化している。本トピックは、生成 AI によるわいせつ物の自動生成とその流通がもたらす社会的影響を扱うものであり、単なる「違法コンテンツ生成」の範疇を超え、倫理的課題や制度的な空白、ガバナンスの不備を浮き彫りにする社会技術的な問題として捉える必要があると考えられる。

特に児童を対象とする AI 生成物は「AI-generated Child Sexual Abuse Material (AI-CSAM)」として国際的に問題視されており、実在する児童か否かを問わず規制の不備が指摘されている。女性に対する被害も顕著であり、SNS 画像の無断利用によるディープフェイクポルノは数千万件規模で流通している。さらに、こうした被害は社会的抗議や法改正運動を引き起こし、国際的な規制強化の流れを生んでいる。

上述のように、生成 AI によるわいせつ物生成・流通は、個人の尊厳や権利を直接侵害するのみならず、社会規範や法制度のあり方全体に影響を与える重要課題と位置づけられると考えられるため、本事業における検討対象としている。

■　文献調査結果

生成 AI を利用したわいせつ物の生成・流通は、国内外で急速に深刻化しており、とりわけ児童や女性への権利侵害と性的被害の側面が顕著である。

まず児童に関しては、未成年者を対象とする性的なわいせつ画像が短時間で生成可能となっており、国内では 2024 年の警察への相談件数がすでに 100 件を超えているとの報道もある [27]。米国でも SNS に投稿された未成年者の画像が生成 AI によってわいせつ画像に変換される事例が急増し、法執行機関が取り締まりを強化している [28]。さらに国際的には、オープンソースの生成 AI と追加学習技術（LoRA）を利用して「実在児童」に酷似した AI-CSAM が大量生成されている実態が調査報道によって明らかにされている [29]。また、大規模学習データセット（例：LAION）に「実在児童」を含むわいせつ画像が混入し

[27] 読売新聞,「卒業アルバム加工した偽の性的画像ＳＮＳ拡散、一部は小中高生らが作成…警察庁がＡＩサイト調査」
https://www.yomiuri.co.jp/national/20250831-OYT1T50010/

[28] Forbes, “Pedophiles Are Using AI To Turn Children’s Social Media Photos Into CSAM”
https://www.forbes.com/sites/thomasbrewster/2025/04/08/pedophiles-use-ai-to-turn-kids-social-media-photos-into-csam/

[29] Pulitzer Center, With AI, “Illegal Forums Are Turning Photos of Children Into Abusive Content”
https://pulitzercenter.org/stories/ai-illegal-forums-are-turning-photos-children-abusive-content

ていたことも確認[30]されており、開発段階でのデータ管理の不備が不適切な生成を助長している。このように、「実在児童」の被害だけでなく、実在しない児童を対象とした「非実在児童」のポルノも拡散しており、後述する日本においては現行法の規制が及ばない領域でのリスクが拡大している。

女性に対する被害も無視できない深刻な問題である。オックスフォード大学の報告によれば、オンラインプラットフォーム上で公開されているディープフェイクモデルは 3 万 5,000 件を超え、ダウンロード数は 1,500 万回以上に達している。その多くが女性の顔写真を無断で性的コンテンツに加工したものであり、プライバシー侵害やリベンジポルノ被害が世界規模で発生している[31]。SNS などから入手した日常的な写真が悪用されるため、対象者は意図せず被害者となり、心理的な負担や社会的ダメージが長期的に残ることが懸念される。

このような状況は、法規制にも大きな影響を及ぼしている。韓国では未成年女性が被害を受けた事件を契機に、2024 年 9 月に性的なディープフェイクの所持・視聴・購入・保存を違法化する法改正が行われた[32]。英国では 2023 年に「オンライン安全法」を制定し、ソーシャルメディアや検索サービスを規制対象として違法コンテンツ対策を義務付けたうえ、域外適用を可能[33]とした。さらに 2025 年には「AI を用いた CSAM 生成ツールの所持・流通禁止」を発表[34]し、違反者に禁錮刑を科すなど、AI を利用した児童ポルノへの規制を一層強化している。一方、日本では全国的な法整備が遅れており、鳥取県が 2024 年に鳥取県青少年健全育成条例の改正を行い、児童ポルノ等の作成、製造、提供を禁止した[35]が、対象は「実在児童」に限定されており、規制の範囲には限界がある。明治大学の教授は「非実在児童」の生成画像は処罰対象外となる可能性があると指摘[36]し、現行法の抜け穴が悪用を助長している現状を示している。

次に、生成 AI によるわいせつ物の作成や流通の影響に関連しうるステークホルダーに

---

[30] The Guardian, “AI image generators trained on pictures of child sexual abuse, study finds”
https://www.theguardian.com/technology/2023/dec/20/ai-image-generators-child-sexual-abuse

[31] The ACM Digital Library, “Deepfakes on Demand: The rise of accessible non-consensual deepfake image generators”
https://dl.acm.org/doi/10.1145/3715275.3732107

[32] The Associated Press, “In South Korea, deepfake porn wrecks women’s lives and deepens gender conflict”
https://apnews.com/article/south-korea-deepfake-porn-women-df98e1a6793a245ac14afe8ec2366101

[33] 西村あさひ法律事務所,「英国オンライン安全法（Online Safety Act）の解説〜その適用範囲と要対応事項の概要〜（2023 年 12 月 13 日号）」
https://www.nishimura.com/ja/knowledge/newsletters/europe_231213

[34] 英国政府, “Britain's leading the way protecting children from online predators”
https://www.gov.uk/government/news/britains-leading-the-way-protecting-children-from-online-predators

[35] 鳥取県,「鳥取県青少年健全育成条例の改正について」
https://www.pref.tottori.lg.jp/320988.htm

[36] 明治大学,「生成 AI によって引き起こされるサイバー犯罪に関する法整備」
https://www.meiji.net/life/vol539_ishii-tetsuya

ついて考察する。本影響に関連しうるステークホルダーとして、例えば、AI 開発者、AI 提供者、エンドユーザーが挙げられる。AI 開発者 は、データセットの選定やモデル設計に直接関与するため、不適切なデータの混入防止や透明性確保が重要である。AI 提供者は出力制御やコンテンツモデレーションを担う立場にあり、違法生成物の流通防止に直結する責任を負うと考えられるという点で関連する。エンドユーザーは加害者と被害者の双方を含み、とりわけ未成年や女性が被害を受けやすい。さらに、政府や立法機関は法的枠組みの整備を担っており、英国や韓国など一部の国では規制強化の動きが見られる一方で、日本では今後も検討すべき課題が残されていると考えられる。

■　インタビュー調査結果

本トピックに関するインタビューは、児童保護団体 E 社、ネットパトロール団体 F 社に対して実施した。E 社は AI-CSAM に関する現状を把握していると想定され、F 社はネットパトロールを行っており AI により生成されたわいせつ物の流通状況を把握していると想定されるため、インタビューの対象組織として選定した。ここでは、インタビューから得られた要旨（回答者の発言）を記載することとし、各インタビューの詳細については A.2 のインタビュー議事を参照いただきたい。

●E 社に対するインタビュー調査において把握した事項

E 社によれば、AI-CSAM の国内統計は、AI 由来か否かの判断が困難であるため、統計的な情報を取得することが難しいとしている。一方、海外ではアメリカの NCMEC（全米行方不明・被搾取児童センター）やイギリスの IWF（インターネット監視財団）の報告では急増が確認され、日本でも相談件数は体感的に増加しているという。議論の最中であるが、CSAM は①完全実写、②部分実写（AI による加工）、③テキスト・音声、④アニメ、⑤その他すべての CSAM、の 5 段階に整理しており、対策優先度は①→⑤の順と考えている。特に「一部実在児童（顔や体だけ実在児童のディープフェイク）」は被害が深刻だが、現行の児童ポルノ禁止法は「実在児童」を前提とするため、非実在児童や一部実在児童には適用が不明確で、法執行機関も運用に苦慮していると考えられる。また、刑法 175 条のわいせつ物頒布罪はあるものの刑が軽く、抑止力不足が課題であるとしている。

対策として E 社はまず、児童ポルノ禁止法の改正または新法制定により、実在する児童に加え、実在する児童へ擬似しているものについても規制対象とすることを関係省庁に提言しているという。さらに、警察内に AI-CSAM 専門部署を設けること、相談窓口の拡充と敷居の引下げ、被害児童の救済スキーム整備、子供への AI・CSAM リテラシー教育、AI に CSAM を学習させず生成を技術的に防止すること等も提言しているという。短期的には自治体条例の活用やプラットフォーム事業者による自主削除システム「TakeItDown」の参加拡大が有効とみている。

国際連携では各国 NGO・政府と情報交換を行っており、課題は上述の通り児童ポルノ

禁止法の定義であるとしている。海外はAI開発者からプラットフォームまで幅広いステークホルダーへの規制を検討している国もあり、E社は「被害者が利用しやすい運用」を優先して国内制度の再構築を行うことが考えられると指摘している。

●F社に対するインタビュー調査において把握した事項

F社は2025年3～6月、閉鎖型コミュニティや匿名掲示板を調査し、実在児童の画像を加工したディープフェイクポルノを250件以上、うち小学生被害を20件確認している。なお、実在児童か否かの画像鑑別は報道機関や研究機関の協力により95～99%の精度を得ている。被害は著名人が最多だが、2023年頃から一般の中高生が急増し、静止画から動画生成へと質も悪化している。生成されたディープフェイクポルノ画像や動画はWebサイトで販売されており、中にはディープフェイクポルノへの加工を請け負う者も存在する。

通報について、被害者の所属がわかるものについては学校、教育委員会、管轄の警察へ、所属が分からないものについてはインターネットホットライン、NCMECへ通報を行っている。しかし、仮に削除された後も着衣元画像が残り再悪用されることを懸念している。また、加害者は被害者の画像を所持している同級生等の近しい人物により作成される場合が多く、被害者は疑心暗鬼になる。さらに、現状の法規制では名誉棄損で開示請求することが難しい場合があり、被害者は民事で解決するしかない状況における被害者の経済的負担も課題視している。

対策として、法規制の観点では実効性ある法整備と国際捜査協力を行うことが望ましいと考えている。技術的な観点では、CSAM生成可能サービスへのアクセス制限、わいせつなデータを含まないデータセットによる学習、わいせつ物の生成制限を行うことが望ましいと考えている。

●E社、F社のインタビュー調査結果を踏まえた考察

上記のE社、F社へのインタビューから、AIによるわいせつ物の生成中でも特にCSAM（AI-CSAM）が問題視されていることが理解できる。また、AI-CSAMの被害実態について、海外では専門機関の報告によると増加しているとのことである。一方、日本ではインタビュー対象組織の体感では増加傾向にあるものの、統計的な情報を取得することが難しいという。そのため、今後国内の被害実態を調査することが重要であると考えられる。さらに、CSAMの生成について、現状の法規制は実在する児童に関するわいせつ物を規制対象としていることから、取り締まりが容易でないことが1つ課題であると考えられる。そのため、今後AIがますます普及することを前提とした法規制の議論を進めていくことや、そもそもわいせつ物を生成させないような出力制御等を行っていくことが重要であると考えられる。

3.2 経済活動への影響

3.2.1 生成物の権利に関する懸念

■ トピックの概要説明

生成 AI は、テキスト・画像・音声・映像など多様な表現を短時間で生成できる利便性から、創作活動やエンターテインメントの分野に大きな革新をもたらしている。しかし、その一方で既存の著作物を無断で学習データとして利用し、キャラクターや作風を模倣することで著作権侵害やブランド価値毀損を招く可能性のある事例が国内外で報告されている。また、人物の顔や声の無断利用によるパブリシティ権・肖像権侵害、さらには「本人性の喪失」といった人格的な被害が生じることもある。加えて、故人を生成 AI で再現するサービスが登場し、遺族の心情ケアといったポジティブな側面がある一方で、人格の歪曲や商業的利用など新たな倫理的リスクを内包している。

上述のように、生成 AI の生成物に関する権利懸念は、著作権・パブリシティ権・肖像権といった既存の法制度の枠組みを超えた社会技術的課題であり、文化的オリジナリティや個人の尊厳を脅かす可能性があるため、本事業における検討対象としている。

■ 文献調査結果

生成 AI の権利に関する懸念として、著作権・文化的価値の毀損、パブリシティ権・肖像権の侵害、そして故人再現に関する新たな倫理的課題を取り上げる。

まず、著作権の側面では、中国の裁判所が AI 生成の「ウルトラマン風画像」に対し著作権侵害を認めた判決[37]や、米国で Walt Disney 社と NBCUniversal 社が Midjourney 社をキャラクター模倣で提訴した事例[38]が象徴的である。これらは生成 AI が権利者の経済的利益を奪い、ブランド価値を毀損する現実を示している。また、「スタジオジブリ風」画像の拡散は文化的盗用として批判され、日本の現行法では作風は保護対象外である[39]一方、米国では作風も保護対象とすべきとの立法論が議論[40]されている。報道・出版業界でも、The New York Times 社が記事の無断学習を理由に OpenAI と Microsoft を提訴する[41]

[37] 読売新聞,「「ウルトラマン」に似た画像提供の生成ＡＩ事業者、中国の裁判所が著作権侵害で賠償命令」
https://www.yomiuri.co.jp/culture/subcul/20240415-OYT1T50069/
[38] 日本経済新聞,「ディズニーなど、米 AI 新興を著作権侵害で訴え　映画大手で初」
https://www.nikkei.com/article/DGXZQOGN11DXI0R10C25A6000000/
[39] 日経 xTECH,「生成された文章や画像　AI の著作権はどうなっているのか、文化庁の見解を知る」
https://xtech.nikkei.com/atcl/nxt/column/18/02737/061600037/
[40] 日本経済新聞,「AI 製のジブリ風画像が世界で流行、「作風」保護の議論再燃」
https://www.nikkei.com/article/DGXZQOGN28CZL0Y5A320C2000000/
[41] 日本経済新聞,「米 NY タイムズ、OpenAI を提訴　記事流用で数千億円損害」
https://www.nikkei.com/article/DGXZQOGN27CXP0X21C23A2000000/

一方、The Washington Post 社は提携により記事利用を許諾する契約モデルを構築する[42]など、メディアと AI 開発者の間で権利を巡る新しい関係性が模索されている。音楽業界でも全米レコード協会が AI 音楽生成サービスを提訴[43]し、日本の JASRAC を含む団体が文化庁に意見を提出する[44]など、著作権保護を求める動きが強まっている。クリエイター個人も深刻な影響を受けており、日本芸能従事者協会の調査では 9 割以上が権利侵害を実感していると回答している[45]。

次に、人物の顔や声の無断利用によるパブリシティ権・肖像権侵害に関する事例を取り上げる。肖像パブリシティ権擁護監視機構の調査によれば、主要 SNS で「〜になってみた」「〜に歌わせてみた」といった投稿が延べ 8 万件以上確認され、総閲覧数は 2.6 億回に達している[46]。他には、歌手 Drake や The Weeknd の声を AI で模倣した楽曲が 1,000 万回以上再生され削除要請が出された事例や俳優トム・ハンクスの偽広告出演[47]などがあり、本人の社会的評価や契約関係に深刻な影響を与えている。日本でも声優や俳優の声が無断で学習データに利用され、労組が「声の肖像権」の法整備を要望する[48]など保護の動きが広がっている。さらに、一般人も被害対象であり、SNS 投稿が無断で学習され、わいせつ物の生成や流通（3.1.4 節）で取り上げたように性的・暴力的コンテンツに登場させられる事例が報告されている。また、こうした無断利用は「本人性の喪失（自分自身のアイデンティティや人格的な同一性が、本人のコントロールを離れて希薄化してしまう概念的な被害）」と呼ばれる人格的被害を生み、自己証明の基盤を揺るがしている。

最後に、生成 AI を用いた故人再現サービスに伴う倫理的懸念について整理する。故人の再現は、遺族へのグリーフケアや歴史的人物の文化資産保存に資する一方で、人格の歪曲や長期的な悲嘆状態への依存を引き起こす危険性が指摘されている。ケンブリッジ大学の研究では、驚くほど正確な AI 再現により、故人から望まない形で影響を受け、人々に大きな精神的な苦痛をもたらすリスクが指摘されている。また、営利目的による故人の商業利用も懸念され、法的にも故人の同意取得や遺族の権限をどのように位置づけるかは未

---

[42] 日本経済新聞,「OpenAI、ワシントン・ポストと提携　検索に記事利用」
https://www.nikkei.com/article/DGXZQOGN22DXC0S5A420C2000000/

[43] 日本経済新聞,「世界音楽大手、生成 AI の新興 2 社提訴　著作権侵害を主張」
https://www.nikkei.com/article/DGXZQOGN250JB0V20C24A6000000/

[44] JASRAC,「「AI と著作権に関する考え方について（素案）」 に関して文化庁へ意見を提出しました」
https://www.jasrac.or.jp/information/release/24/02_3.html

[45] 一般社団法人日本芸能従事者協会,「全クリエイター実態調査アンケート 10　AI リテラシー(集計結果)」
https://artsworkers.jp/questionnaire/20230608/

[46] 特定非営利活動法人 肖像パブリシティ権擁護監視機構,「生成 AI 時代における肖像権・パブリシティ権等における侵害疑義事案の実態を初調査 〜業界初の大規模実態調査で判明 現状と今後の課題が浮き彫りに〜」
http://www.japrpo.or.jp/img/pressrelease20250624.pdf

[47] 実演家著作隣接権センター,「生成 AI と実演―パブリシティ権を巡るアメリカの動向―」
https://www.cpra.jp/cpra_article/article/000762.html

[48] 日本俳優連合,「生成系 AI 技術の活用に関する提言」
https://www.nippairen.com/about/post-14576.html

解決の課題とされている[49]。さらに、コーネル大学の研究では、生成AIが故人の情報を基に新しいコンテンツを生み出す一方で、データの出典や文脈が失われ、故人とは異なる発言を生成するリスクが増大することが指摘されている[50]。

次に、生成AIによる生成物の権利に関する懸念で影響に関連しうるステークホルダーについて考察する。本影響に関連しうるステークホルダーとして、例えば、AI開発者、クリエイター・権利者、政府・規制機関が挙げられる。AI開発者は権利に紐づいている学習データの選定を通じた責任を負う点で直接関わる。クリエイターや権利者は収益機会の喪失と職業的地位の低下に直面し、エンドユーザーも無意識のうちに侵害に加担するリスクを抱える。さらに政府・規制機関は、文化庁がガイドラインを発表[51]し、日本新聞協会が声明を出す[52,53]など制度整備を進めているが、国際的な規制調和も課題であると考えられる。

[49] Hollanek, Tomasz, and Katarzyna Nowaczyk-Basińska. "Griefbots, deadbots, postmortem avatars: On responsible applications of generative AI in the digital afterlife industry." Philosophy & Technology 37.2 (2024): 63.
https://link.springer.com/article/10.1007/s13347-024-00744-w

[50] Morris, Meredith Ringel, and Jed R. Brubaker. "Generative ghosts: Anticipating benefits and risks of AI afterlives" Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025.
https://dl.acm.org/doi/10.1145/3706598.3713758

[51] 文化庁,「AIと著作権に関するチェックリスト&ガイダンス」
https://www.bunka.go.jp/seisaku/chosakuken/pdf/94097701_01.pdf

[52] 一般社団法人 日本新聞協会, 生成AIにおける報道コンテンツの無断利用等に関する声明
https://www.pressnet.or.jp/statement/broadcasting/240717_15523.html

[53] 一般社団法人 日本新聞協会, 生成AIにおける報道コンテンツの保護に関する声明
https://www.pressnet.or.jp/statement/broadcasting/250604_15900.html

3.2.2 労働環境への影響

■ トピックの概要説明

生成 AI が技術的に発展し、入出力可能なデータの種類が増加したり、生成物の精度が飛躍的に向上したりしたことで、社会に生成 AI が普及している。そのため、多くの企業では様々な業務へ生成 AI を適用することで業務効率化を図っており、社会全体の労働環境に変化をもたらしている。生成 AI の普及は上述した業務効率化のような労働環境に対するポジティブな影響を及ぼす一方、生成 AI が労働そのものを代替することで、雇用喪失の発生といったネガティブな影響を及ぼす可能性があることが、国内外の研究や事例から指摘されている。

上述のように、生成 AI の普及が労働環境へ及ぼす影響はポジティブな影響、ネガティブな影響の両方の側面があると考えられる。そこで、事例や関連する研究から、企業において生成 AI の利活用がどの程度進んでおり、社会の労働環境がどのように変化しているのかの現状を把握することは重要であると考える。また、雇用喪失の発生といった社会に大きな影響を与え得る事象に対する対策としてどのようなものがあるかを整理することも重要であると考えるため、本事業における検討対象としている。

■ 文献調査結果

生成 AI が労働環境に与える影響について、業務効率化や新規雇用創出といったポジティブな影響に加え、生成 AI が既存の労働を代替することで失業者が発生するといったネガティブな影響について、関連する事例を紹介した上で、研究結果を紹介する。最後に、本影響を受ける可能性のあるステークホルダーを考察する。

まず、本影響に関連する事例を紹介する。日鉄ソリューションズ株式会社によると、生成 AI により翻訳や表計算ソフトの関数作成といった間接業務を効率化し、生成 AI 導入後 3 カ月で 9500 時間を超える業務削減効果が得られた事例があるとされている [54]。また、シンガポールの銀行である DBS は、AI が人間に取って代わって仕事をするとして、2028 年までに従業員 4,000 人を削減する見込みだと発表している [55]。このように、生成 AI の活用が労働環境に様々な影響を与えていることが理解できる。

次に、本影響に関連する研究結果を紹介する。内閣府やハーバードビジネススクールが公開しているレポートによると、生成 AI が労働環境に与える影響を検討する際には以下 2

[54] 日鉄ソリューション株式会社,「生成 AI で間接業務を効率化 導入 3 カ月で 9500 時間の削減に」 https://www.nssol.nipponsteel.com/casestudy/02908.html

[55]BCC NEWS JAPAN,「AI 活用で従業員 4000 人削減へ　アジアの主要銀行 DBS」 https://www.bbc.com/japanese/articles/c8x4qlydnkzo

つの側面を意識することが重要であるとされている[56,57]。1つ目は、生成AIが人間の職業・タスクを完全に置き換え、人間が介在する余地をなくしてしまうような「代替型」としての側面である。従来人間が行ってきたタスクのうち、事務的タスクの多くは、コンピューターの性能の上昇に応じて労力の削減が可能となってきている。今後は、生成AIを活用することで、事務的タスクは更なる効率化が可能となり、部分的にはほぼ完全な自動化まで実現される可能性があると考えられている。このように、人手がかからなくなったタスクについては、結果としてAIが労働者を代替した形となる。

2つ目は、人間の労働を補助して楽にし、生産性を上げ、新たな仕事を生み出すきっかけとなるような「補完型」としての側面である。マサチューセッツ工科大学（MIT）の研究者による実験では、AIの使用により、レポートや電子メールなどのタスクについて生産性が向上されたという結果が得られており、AIは人のタスクや職業を「補完する」機能があるとされる[58]。つまり、労働者の一部のタスクをAIが担い、AIと人間とが協働することとなる。

代替型の側面が強い職業が多い業界の場合、業務効率化によるコスト削減といったポジティブな影響が大きい一方、当該職種の従業員が失業するといったネガティブな影響も大きい可能性があると考えられる。また、補完性の高い職業においても、一部のタスクは効率化・自動化され人手がかからなくなるため、雇用は一定程度減少し得ると考えられている。しかしながら、自動車の発明により整備士などの新たな職業が生まれたように、新たな雇用を生み出す可能性もあるとされる。例えば、東京大学松尾研究室のレポートによると、新たな職業としてプロンプトエンジニアに注目が集まっており、米国のAnthropic社ではプロンプトエンジニアを募集するなど需要が高まっている状況にある[59]。

次に、生成AIの普及により労働環境の変化の影響に関連しうるステークホルダーを考察する。本影響に関連しうるステークホルダーとして、例えば、AI開発者、AI提供者、AI利用者・エンドユーザーが挙げられる。AI開発者は、多くの企業がどのような業務にAIを活用可能であるかはAIモデルの性能に大きく依存すると考えられるという点で関連がある。AI提供者はAIシステムをアプリケーション、製品、既存のシステム、ビジネスプロセス等に組み込むため、AIシステムと企業内の業務との統合を行う役割を持つという点で関連するステークホルダーであると考えられる。

AI利用者・エンドユーザーは、特に情報通信業や金融業・保険業、教育、学習支援業な

---

[56] 内閣府,「世界経済の潮流 2024年I」
https://www5.cao.go.jp/j-j/sekai_chouryuu/sh24-01/s1_24_1_1.html

[57] Harvard Business School, "Displacement or Complementarity? The Labor Market Impact of Generative AI"
https://www.hbs.edu/ris/Publication%20Files/25-039_05fbec84-1f23-459b-8410-e3cd7ab6c88a.pdf

[58] 内閣府,「新しい資本主義のグランドデザイン及び実行計画2023改訂版案」
https://www5.cao.go.jp/keizai-shimon//kaigi/minutes/2023/0616/shiryo_01-3.pdf

[59] 松尾研究所,「生成AI時代の人材育成」
https://www.meti.go.jp/shingikai/mono_info_service/digital_jinzai/pdf/008_05_00.pdf

どの情報やデータを扱うオフィスワークが中心の産業が特に関連すると考えられる。これは、国内外で産業や職業別に労働補完率を推計し労働の代替性を評価する調査研究がなされており、上述したオフィスワーク中心の産業の業務は AI の得意領域であり、労働の代替性が高いとされているためである [60,61]。

■ インタビュー調査結果

本トピックに関するインタビューは、金融サービス業 G 社、IT サービス業 H 社に対して実施した。G 社、H 社ともに生成 AI を業務に積極的に利用していることから、インタビューの対象組織として選定した。ここでは、インタビューから得られた要旨（回答者の発言）を記載することとし、各インタビューの詳細については A.2 のインタビュー議事を参照いただきたい。

●G 社に対するインタビュー調査において把握した事項

G 社では、社内ではメール作成・翻訳・要約・WEB 検索等の一般業務や RAG（検索拡張生成）、代理店でも一般業務に加え、RAG を活用したマニュアル・約款・商品情報の回答生成を行っている。代理店は営業だけでなく契約保全の問い合わせやお礼状作成、営業話法の練習にも生成 AI を利用している。社内累計利用率（1 回以上 AI を利用した社員の割合）は 80%以上でありアンケートベースではあるものの効率化は 30%程度できているという声がある。また、代理店からは 70%以上から継続して利用したいとの声を得ている。

顧客向けの AI システムについては、生成 AI を利用したアバターにより顧客からの問い合わせ回答を行う AI システムを先行リリースしている。前提として、G 社では AI ガバナンス強化の下、リスクを 4 段階で管理しており、顧客向けの AI システムは最もレベルの高いリスク分類としている。特にハルシネーション対策として日次モニタリングとハルシネーションが発生し、顧客に誤認を与えるような場合は訂正連絡を行うことを検討しているが、日次モニタリングの負荷が高いことが課題であるとしている。

また、社員は先輩や WEB 検索を行う前にまず AI へ質問する習慣が定着し、資料作成やアイデア創出等の業務の効率化と品質向上がなされている。今後 AI エージェントの活用が進むと、AI の検証、監視といったマネジメントや、活用方法を考える業務にシフトすると考えられる。通常の業務、金融サービスに関する業務、人間が実施していた業務については、AI エージェントで実施することができると考えている。ただ、お客様の心情に寄り添う業務などは、人間の仕事として残るものであると思われる。また、生成 AI 導入にあたって、月 1 回の研修と部署別ワークショップで社員への教育を推進している。

---

[60] 大和総研,「生成 AI が日本の労働市場に与える影響②」
https://www.dir.co.jp/report/research/economics/japan/20231211_024139.pdf

[61] International Labour Organization, "Generative AI and Jobs"
https://webapps.ilo.org/static/english/intserv/working-papers/wp140/index.html#ID0E4C

●H 社に対するインタビュー調査において把握した事項

H 社では、エンジニアはコード補助 AI ツールを複数導入し、AI の利用が前提となっている。どの部署でも何らかの形で AI を活用していると考えられ、例えばセールスやマーケティングでは商談書き起こしや顧客情報検索等で AI を活用している。また、従業員の定型問合せ対応のチャットボットに AI を活用しており、従業員のフィードバックを基に継続的にチューニングしている。

AI 導入により業務の簡略化・ミス削減・高速化だけでなく、プレゼン練習時に的確なアドバイスやモチベーションを引き上げるようなフィードバックといった情緒的価値も創出している。また、上司や同僚に依存せず高度なフィードバックが即座に得られるため、生産性が向上している。リスク面では、情報入力範囲のルールが設けられているが、AI 活用を早く進めるため、一部の特にリテラシーの高い社員に対しては承認の基準を下げ対応することもある。また、AI による業務代替への不安は少なく、代替された場合は新たな仕事へシフトできるという認識が浸透している。

また、全社的な AI リテラシー向上のため、啓蒙活動と研修を強化する方針である。部署単位での AI 活用は自主的に進められ、特にエンジニア部門は積極的に推進している。将来的には AI と共に働くことを前提とした業務・組織設計が必要となり、業務ベースの配置から AI 活用を前提とした配置へ転換することもあり得る。

●G 社、H 社のインタビュー調査結果を踏まえた考察

上記の G 社、H 社へのインタビューの結果から、デスクワークの様々な業務に対して AI が活用されており、人間の業務が効率化されていることが理解できる。つまり、現時点では AI の活用により人間の業務を「補完」する側面が強いと考えられる。また、業務効率化の側面だけでなく、人間のモチベーションを向上させるような情緒的な価値ももたらすことが考えられる。一方で、今後さらに AI の性能が向上することで、これまで人間が実施してきた業務を AI が代替する可能性もあると考えられる。そのため、AI の利用を前提として、従業員の AI リテラシーの向上や業務の再設計に取り組むことが考えられる。また、AI の利活用を促進していくために、ハルシネーションといった AI 固有のリスクに対しても適切に対処していくことも重要であると考えられる。

3.2.3 生成物の氾濫が及ぼす影響

■ トピックの概要説明

生成 AI は、低コストかつ短時間で大量のメッセージやコンテンツを作成することを可能にしている。ChatGPT 等の生成 AI ツールにより、プロンプト一つで人間が作成したものと同じような文章や画像が瞬時に生成でき、特に言い換え・パラフレーズ機能は同一内容を無数のバリエーションで表現できる。この特性は、アンケートやレビュー、SNS 投稿といった多様な場面で従来のスパム検知機能を突破し、質より量を優先した投稿を可能にしている。こうして氾濫する生成物は、エンドユーザーが求める有益な情報とは対極の、大量で粗悪な情報であり、ランキングアルゴリズムによって上位表示されることで、優良コンテンツの可視性を奪う事態が報告されている。その結果、プラットフォームは従来のパターンマッチ型判定から、生成 AI を活用した新たなスパム検知機能への移行を余儀なくされている。このように、生成 AI による生成物の氾濫は、社会に大きな影響を与える可能性があり、重要な課題であると考えられるため本事業における検討対象としている。

一方、上述のようなネガティブな影響だけでなく、従来から存在した偽レビュー等に対する規制整備の加速といったポジティブな影響ももたらしている。そこで以下では、生成 AI による生成物の氾濫が及ぼす影響について、ネガティブな側面のみならず、ポジティブな側面についても整理している。

■ 文献調査結果

生成 AI による生成物の氾濫は、プラットフォームの想定を超える規模で生じている。その影響は調査・レビューの信頼性低下、ソーシャルメディアの汚染、検索サービスの品質劣化といったネガティブ側面に加え、規制整備の加速といったポジティブ側面も確認されている。

まず、ネガティブな影響については 3 つの事例を取り上げる。第 1 に、調査・レビューの信頼性の悪化である。スタンフォード大学の Janet Xu 准教授によれば、オンライン調査の参加者の約 3 分の 1 が ChatGPT 等の AI ツールを用いて回答を作成していることが明らかになった [62]。その結果、データ品質が均質化し、例えば「返信には誤字が少なかった、しかも不自然に丁寧だった。」、「LLM は一貫してより中立的で抽象的な言語を使用しており、人種や政治などのデリケートな話題に対してより距離を置いて接している可能性があることを示唆している」といった特徴が確認された [63]。このようなデータは研究や政策形

[62] Zhang, Simone, Janet Xu, and A. Alvero. "Generative ai meets open-ended survey responses: Participant use of ai and homogenization".
https://www.gsb.stanford.edu/faculty-research/working-papers/generative-ai-meets-open-ended-survey-responses-participant-use-ai

[63] Stanford Report, “AI-generated survey responses could make research less accurate – and a lot less interesting”
https://www.gsb.stanford.edu/insights/ai-generated-survey-responses-could-make-research-less-accurate-lot-less-interesting

成における基礎資料の信頼性を損なう可能性がある。第2に、ソーシャルメディアエコシステムの汚染である。Harvard Kennedy School と Stanford 大学の共同研究では、生成 AI により作成された画像を利用したスパムページが Facebook で急速に拡散し、単一投稿で数百万のエンゲージメントを獲得した事例が報告されている[64]。また、発展途上国のクリエイターが米国市場向けに高額の広告収入を狙い、生成 AI によってコンテンツを大量生産している実態もメディア記事で明らかになっている[65]。エンドユーザーはそれを生成 AI により作成されたものであると気づかずに消費するため、プラットフォーム全体の情報環境が汚染される。第3に、検索サービスの品質劣化である。Google 検索において低品質コンテンツの比率が上昇し、検索結果上位が AI 生成コンテンツに占められる事態が確認されている。Google は AI 生成コンテンツが検索結果の上位を占める状況に対して、年に数回の「スパム対策アップデート」を実施している[66]。検索の信頼性が低下すれば、インターネット全体の情報利用基盤に影響を及ぼす可能性がある。

一方でポジティブな影響として、生成 AI の氾濫は規制強化を促す契機にもなっている。偽レビューは生成 AI 普及以前から存在していたが規模が限定的であったため、厳格な規制は存在しなかった。しかし生成 AI によって作成されたレビューの投稿が大規模化し、消費者被害や市場歪曲の影響が拡大したことから、米連邦取引委員会（FTC）は 2024 年に生成 AI によって作成されたレビューを含む偽レビューを規制対象とし、違反時には民事制裁を科す方針を明確化した[67]。これは法制度が生成 AI のリスクに対応し始めた重要な一歩と評価できると考えられる。

次に、生成 AI による生成物の氾濫が及ぼす影響に関連しうるステークホルダーについて考察する。本影響に関連しうるステークホルダーとして、例えば、AI 開発者、AI 提供者、エンドユーザーが挙げられる。AI 開発者は、悪意あるエンドユーザーにスパム生成を可能にするツールを提供する一方で、プラットフォームのスパム検知を支援する技術開発の担い手でもある。つまり、加害と防御の両側面に関与し得る立場にあると考えられる。AI 提供者の中でも特に自社のプラットフォームに AI 機能を組み込んだプラットフォーマーは本影響に関連する主要なステークホルダーであると想定される。特に検索エンジン、

---

[64] Harvard Kennedy School Misinformation Review, “How spammers and scammers leverage AI-generated images on Facebook for audience growth”
https://misinforeview.hks.harvard.edu/wp-content/uploads/2024/08/diresta_spammers_scammers_ai_images_facebook_20240815.pdf

[65] NRP, “AI-generated spam is starting to fill social media. Here’s why”
https://www.npr.org/2024/05/14/1251072726/ai-spam-images-facebook-linkedin-threads-meta

[66] Search Engine Roundtable, “Google August 2025 Spam Update Unleashed”
https://www.seroundtable.com/google-august-2025-spam-update-40008.html

[67] Federal Trade Commission (FTC), “Federal Trade Commission Announces Final Rule Banning Fake Reviews and Testimonials”
https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials

SNS、レビューサイト、アンケートプラットフォームといったランキング・インセンティブ機能を有するサービスは、生成 AI によるスパム投稿に直接晒されやすいと考えられる。結果として、ユーザー体験の劣化や広告収益への悪影響が生じ、プラットフォーム運営の持続可能性を脅かす可能性がある。エンドユーザーの中のスパム投稿を行う者は、匿名性や即時性に支えられ、リワードや収益を目的に生成 AI を濫用すると想定される。こうした動機は個人から組織に至るまで幅広く存在し、利用が容易であるほど悪用の誘因が高まると考えられる。

3.2.4 経済格差へ及ぼす影響

■　トピックの概要説明

近年、生成 AI の性能が飛躍的に向上し、API などを通じて誰もが比較的容易に利用できるようになったことで、社会経済活動のあらゆる側面に変革をもたらす汎用技術としての能力が注目されている。

これまでの AI が主に「分析」や「判断」を担ってきたのに対し、生成 AI は「創造」の領域に踏み込んだ点に本質的な違いがある [68]。これにより、従来は人間にしかできないとされてきた創造的なタスクの一部が自動化の対象となり、生産性の劇的な向上が期待される一方で、後述するような個人の資本所得、企業の市場評価などに変化を及ぼす可能性が指摘されている。

特に経済格差の文脈では、生成 AI がもたらす恩恵が社会全体に均等に行き渡らず、特定の人々や企業、国家に富が集中することで、既存の格差をさらに拡大、あるいは新たな格差を生み出すリスクが懸念されている。以上より、本影響についての現状や対策を把握することは重要であると考えられるため、本事業における検討対象としている。

■　文献調査結果

生成 AI の普及は、経済格差に多方面で影響を及ぼす可能性がある。ここでは、個人（職種）、企業、国の 3 つの観点で本影響を取り上げる。

まず、個人（職種）への影響として、労働環境への影響（3.2.2 節）で記載した通り、生成 AI は事務職やコールセンター業務、文章やコード生成といったルーティン作業を中心に代替性が高いことが指摘されており、本節では経済格差の観点で論じる。
International Monetary Fund （IMF）の調査によれば、労働所得への影響は生成 AI との補完性に左右され、生成 AI と補完的なスキルを持つ労働者は所得が増加する一方で、代替されやすい職種では所得の伸びが小さい。その結果、全体の生産性が向上しても、所得増の恩恵が偏り、労働所得格差が拡大する懸念がある。また、生成 AI への投資や資産を持つ層に利益が集中することで、資本所得や資産格差も拡大しやすいとされる [69]。一方、マサチューセッツ工科大学（MIT）は、生成 AI が高所得層に求められる管理業務や高度な開発業務を代替する場合、これまで優位にあった高所得者層の労働価値が下がり、格差緩和につながる可能性もあると指摘している [70]。

企業への影響については、AI 開発に必要な膨大な計算資源やデータ、人材を確保でき

[68] Gartner,「生成 AI とは何？仕組みや従来の AI との違い、活用例、注意点をわかりやすく解説」
https://www.gartner.co.jp/ja/topics/generative-ai
[69] International Monetary Fund (IMF), “Gen-AI: Artificial Intelligence and the Future of Work “
https://www.elibrary.imf.org/view/journals/006/2024/001/article-A001-en.xml
[70] MIT Press, “Generative AI and the Future of Inequality“
https://mit-genai.pubpub.org/pub/24gsgdjx/release/1

るのは一部の海外大手テクノロジー企業に限られており、彼らが提供するプラットフォームへの依存が進めば企業間格差はさらに広がると考えられる。生成 AI を事業に活用する上でも、大企業と中小企業との間で導入・活用レベルに差が生じ、生産性や競争力の格差、すなわち企業間格差が拡大するリスクがある。日本国内においては、大企業の半数以上が生成 AI 活用方針を持つのに対し、中小企業は 3 割程度[71]にとどまり、この差が将来の生産性や競争力の差として表れる可能性があると考えられる。また、米国の経済研究機関のレポートによれば、生成 AI の進展により、テクノロジーに特化した企業と伝統的な企業の間の株価のパフォーマンスの乖離が加速していることが示されており、市場評価においても格差が拡大している[72]。こうした流れは、資金力やデータ基盤を持つ企業がさらに優位に立ち、業績格差を固定化する要因となる。

国への影響では、デジタル赤字が大きな課題である。デジタル赤字とは、広告料やソフトウェア利用料、クラウドサービス費用、著作権使用料などを含むデジタル関連収支が赤字となる状況を指す。日本では 2024 年に既に約 6 兆円の赤字となっており、AI 関連サービスの多くを海外企業が担っている現状を踏まえると、2035 年には約 28 兆円に達するとの予測もある[73]。つまり、生成 AI の普及が進むほど海外への依存度が高まり、所得の国外流出や国際競争力の低下を招く可能性がある。

次に、生成 AI による経済格差へ及ぼす影響に関連しうるステークホルダーについて考察する。本影響に関連しうるステークホルダーとして、例えば、AI 開発者・AI 提供者、AI 利用者・エンドユーザー、教育・訓練機関、政府・規制当局が挙げられる。AI 開発者は、エンドユーザーや企業が支払う利用料の多くが収益となるという点で関連があると考えられる。AI 提供者は、生成 AI をアプリケーションやビジネスプロセスに組み込み、経済価値を創出する存在として関連があると考えられる。具体的には、社内ツールを提供するプラットフォーマーやシステムに統合して提供するシステムインテグレーターが該当する。AI 利用者・エンドユーザーは、特に情報通信業や金融・保険業、教育・学習支援業など情報を扱うオフィスワーク中心の産業で影響が大きいと考えられる。政府・規制当局は、生成 AI による経済格差への対応、国民の AI リテラシー向上、開発・利活用方針の策定など、多角的な政策対応が求められると考えられる。

---

[71] 総務省,「令和 7 年版情報通信白書（概要）」
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r07/summary/summary01.pdf

[72] National Bureau of Economic Research, “Generative AI and Firm Values“
https://www.nber.org/system/files/working_papers/w31222/w31222.pdf

[73] 経済産業省,「デジタル経済レポート」
https://www.meti.go.jp/policy/it_policy/statistics/digital_economy_report/digital_economy_report.pdf

3.2.5 機密情報の学習や漏洩の懸念

■ トピックの概要説明

近年、生成 AI は飛躍的に進化し、テキスト生成、画像合成、音声合成など多岐にわたる分野で利用が広がり、社会に大きな変革をもたらしている。しかし、その一方で、生成 AI による機密情報の学習や漏洩のリスクが国内外で指摘されている。生成 AI は大量のデータを学習する過程で、機密情報や個人情報を含むデータを取り込む可能性があり、またエンドユーザーが入力した情報が再利用される場合もある。その結果、生成過程で機密情報が外部に出力され、プライバシー侵害や企業の競争力低下、さらには法的リスクを引き起こす恐れがある。つまり、生成 AI による機密情報の学習や漏洩の懸念はエンドユーザー個人の問題だけでなく、企業、行政、社会全体の信頼を損なう可能性がある。そのため、機密情報の管理と保護は社会全体で取り組むべき重要課題であると考えられるため、本事業における検討対象としている。

■ 文献調査結果

生成 AI による機密情報の学習や漏洩がもたらす影響は多面的であり、ここでは「企業の競争力と信頼の棄損」「個人のプライバシー侵害」「法的・コンプライアンスリスクの増大」の 3 点で整理する。

第 1 に、企業の競争力と信頼の棄損について記載する。2023 年、韓国サムスン電子の社員が ChatGPT に社内機密のソースコードを入力した事例が報告された。入力情報は外部サーバーに保存され削除困難となり、他ユーザーに開示される可能性が懸念されたため、同社は生成 AI 利用を原則禁止する新たなポリシーを策定した [74]。ChatGPT がデフォルトでチャットログを保存し学習に利用する仕様であったこともリスクを高めた要因とされる。この事例は、利便性の裏側で企業の知的財産や戦略が流出する危険を示している。

第 2 に、個人のプライバシー侵害の側面について記載する。2023 年、OpenAI 社においてシステム障害が発生し、約 9 時間にわたりユーザーのメールアドレス、支払い先住所、クレジットカード番号の下 4 桁や有効期限などが他ユーザーから閲覧可能となった [75]。不具合の原因はオープンソースライブラリの欠陥とされ、影響を受けたユーザーには通知が行われたが、利用者の信頼を大きく揺るがす事態となった。広範に利用される生成 AI で個人情報の漏洩が起これば、被害は瞬時に拡散し、社会全体の不安を助長する。

第 3 に、法的・コンプライアンスリスクについて記載する。2023 年、米国では ChatGPT がインターネット上の個人情報を違法に収集・学習しているとして、OpenAI 社

[74] Bloomberg,「サムスン、従業員の生成ＡＩ利用を禁止－ChatGPT 経由でデータ漏れる」
https://www.bloomberg.co.jp/news/articles/2023-05-02/RU0AD6T0AFB401

[75] OpenAI, “March 20 ChatGPT outage: Here’s what happened"
https://openai.com/index/march-20-chatgpt-outage/

と Microsoft 社が提訴された[76]。2024 年にはカリフォルニア州地裁が訴状不備を理由に棄却した[77]が、このような訴訟の存在自体が AI 開発者にとって大きな法的リスクを示している。欧州の GDPR など厳格な個人情報保護制度下では、違反すれば巨額の制裁や事業停止に繋がる可能性が高い。

これらの事例を踏まえると、機密情報の学習や漏洩は、企業経営、個人生活、国家安全保障にまで波及する社会的問題であることが分かる。また、国内の状況においては、一般財団法人日本情報経済社会推進協会（JIPDEC）と調査会社の ITR による国内企業の調査によると、生成 AI を使用していくうえでの懸念点として「社内の機密情報を学習データとして利用され情報漏洩すること」が最多であるとされている[78]。また、OWASP の「Top 10 for LLM Applications 2025」では、重大リスクの 1 つとして「Sensitive Information Disclosure（センシティブ情報の漏洩)」が挙げられている[79]。これらは、生成 AI による機密情報の学習・漏洩が社会的に広く認知され、対策が急務であることを裏付けていると考えられる。

次に、生成 AI による機密情報の学習や漏洩の影響に関連しうるステークホルダーについて考察する。本影響に関連しうるステークホルダーとして、例えば、AI 開発者、AI 提供者、AI 利用者・エンドユーザーが挙げられる。AI 開発者は、モデルの設計・学習過程において、入力データが学習されるか否か、また保存・利用の方法を決定する立場にあると想定される。AI 提供者は、生成 AI をアプリケーションや業務プロセスに組み込む立場であり、利用規約や出力制御、アクセス権限管理、コンテンツモデレーションを通じて情報漏洩防止に直接関わると想定される。AI 利用者・エンドユーザーも重要なステークホルダーである。特に国家機密や重要インフラを扱う産業、金融業や医療業界などは特に重要情報を取り扱っているため影響が大きいステークホルダーであると考えられる。

---

[76] Bloomberg,「ChatGPT のオープンＡＩを匿名グループが提訴－個人データ窃取と主張」
https://www.bloomberg.co.jp/news/articles/2023-06-29/RWZS42T1UM0W01

[77] Reuters,「個人情報巡るオープンＡＩ・ＭＳ集団訴訟、米地裁が訴え退ける」
https://jp.reuters.com/economy/industry/RD4KMR7BOFK2NHXYTXMCEE2G7I-2024-05-27/

[78] 株式会社アイ・ティ・アール,「企業 IT 利活用動向調査 2024」
https://www.itr.co.jp/topics/pr-20240315-1

[79] Open Worldwide Application Security Project (OWASP), "OWASP Top 10 for LLM Applications 2025"
https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

3.3 情報空間への影響

3.3.1 偽誤情報の生成や拡散

■ トピックの概要説明

生成 AI は、低コストかつ高速で文章・画像・音声・動画の生成を可能にしており、アイデア出しや文章の要約等、多様な用途で社会に普及している。一方、このような生成 AI の特性を悪用することで、従前は高度な編集・制作リソースが必要だった偽記事や偽ニュース動画が、短時間に作成できるようになっている。また、SNS の普及により、こうした偽情報が容易に拡散され、社会に混乱を与えうる。また、アルゴリズムによる推薦やコミュニティの同質化が重なることで、反証的な情報に接触する機会が乏しくなり、自身の先入観や意見を支持する情報のみを集める傾向が強まり、いわゆるエコーチェンバー現象が形成されうる。なお、「エコーチェンバー現象」とは、価値観の近いエンドユーザー同士が共感を強めることで、特定の意見や思想が過度に増幅され、影響力を持つようになる現象[80]である。さらに、業務で共有される要約・議事録・報告書など「重要文書」にも生成物が混入し、意図せざる誤情報が意思決定へ影響するリスクが高まっている。

上述のように、生成 AI による偽誤情報の生成や拡散は、生成 AI の技術特性と情報社会の構造が相互作用して増幅される社会技術的な課題であるため、本事業における検討対象としている。

■ 文献調査結果

ここでは、記事・動画による偽誤情報の拡散、生成 AI によるエコーチェンバー現象の助長、重要文書への偽誤情報の混入という 3 つの観点で本影響を取り上げる。

まず、生成 AI を悪用した記事・動画による偽誤情報の拡散は、選挙や市場、災害対応など社会の根幹に深刻な影響を及ぼしている。日本経済新聞の記事では、2024 年以降に少なくとも 8 カ国・地域に事例が確認され、台湾や日本の選挙でも候補者や首相を装った偽動画が拡散したと報じられている[81]。2023 年には中国の上場企業・科大訊飛の株価が偽情報で一時 9%急落し、市場混乱を引き起こした[82]。災害時には静岡県台風 15 号の際に AI で生成された偽画像が拡散し、避難や救援を妨害した事例[83]もある。また、SNS はエンゲージメント機能があることからも分かるように、クリック数や反応率の高い記事を優先表

[80] コトバンク,「エコーチェンバー現象」
https://kotobank.jp/dictionary/daijisen/4093/
[81] 日本経済新聞,「選挙イヤーに生成 AI の影　偽画像・音声改変、世界で横行」
https://www.nikkei.com/article/DGXZQOUE186DR0Y4A111C2000000/
[82] 東洋経済新報社,「生成 AI の「偽リスク情報」で中国企業の株価急落」
https://toyokeizai.net/articles/-/676141
[83] 日本経済新聞,「災害時の SNS デマ、警察が厳格姿勢　安易な拡散は要注意」
https://www.nikkei.com/article/DGXZQOUE242390U4A720C2000000/

示される傾向があり、社会的インパクトの大きい偽記事が真実の記事よりも目に触れやすくなる構造的問題があると考えられる。Massachusetts Institute of Technology（MIT）の研究では、偽記事は真実の記事より約 70%拡散されやすく、1500 人に届くまでの時間は 6 分の 1 であるとされ、SNS の構造的要因が拡散を助長している [84]。

2 点目の生成 AI によるエコーチェンバー現象の助長に関して、生成 AI は対話型特性と高度なパーソナライズ機能により、従来以上に情報の偏りを強める危険がある。アメリカの研究 [85]では選択的接触や意見の極性化が短時間で進み、是正策も効果が薄いことが示されている。総務省の「令和 6 年版　情報通信白書 [86]」やデジタル庁が公表した「行政の進化と革新のための生成 AI の調達・利活用に係るガイドライン [87]」では、生成 AI によるエコーチェンバー現象の助長リスクを警告している。そのため、個人レベルでは意見の固定化と認知バイアスの強化、社会レベルでは分断と公共的対話の縮小、政府レベルでは行政や政策形成の偏りといった多層的な影響を及ぼす可能性がある。

3 点目の重要文書への偽誤情報の混入に関しては、社会全体の信頼基盤に深刻な影響を与えうる。学術分野では生成 AI による偽論文が流通している事例 [88,89]や、司法分野では生成 AI による実在しない判例の誤引用 [90]が発生している。企業においても財務資料や契約書において、生成 AI による誤記が経営判断や株価に直結する恐れがある。特に医療、公共安全、環境政策など命や生活に直結する分野では、偽誤情報の混入は社会的パニックや誤った行動を誘発する可能性がある。

次に、生成 AI による偽誤情報の生成や拡散に関連しうるステークホルダーについて考察する。本影響に関連しうるステークホルダーとして、例えば、AI 開発者、AI 提供者・SNS プラットフォーマー、学術機関・報道機関、行政機関、エンドユーザーが挙げられ

---

[84] MIT News, “Study: On Twitter, false news travels faster than true stories”
https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

[85] Sharma, Nikhil, Q. Vera Liao, and Ziang Xiao. "Generative echo chamber? effect of llm-powered search systems on diverse information seeking." Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024.
https://dl.acm.org/doi/pdf/10.1145/3613904.3642459

[86] 総務省,「令和 6 年版　情報通信白書」
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/html/nb000000.html

[87] デジタル庁,「行政の進化と革新のための生成 AI の調達・利活用に係るガイドライン」
https://www.digital.go.jp/assets/contents/node/basic_page/field_ref_resources/e2a06143-ed29-4f1d-9c31-0f06fca67afc/80419aea/20250527_resources_standard_guidelines_guideline_01.pdf

[88] Haider, Jutta, et al. "GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation." Harvard Kennedy School Misinformation Review 5.5 (2024).
https://misinforeview.hks.harvard.edu/wp-content/uploads/2024/09/haider_gpt_fabricated_scientific_papers_20240903.pdf

[89] 読売新聞,「生成ＡＩで日本人の研究者かたり論文捏造か、収入目的の海外サイト「ハゲタカジャーナル」に掲載」
https://www.yomiuri.co.jp/national/20241120-OYT1T50136/

[90] 日本経済新聞,「ChatGPT で資料作成、実在しない判例引用　米国の弁護士」
https://www.nikkei.com/article/DGXZQOGN30E450Q3A530C2000000/

る。AI 開発者は、学習データに含まれる偽誤情報が出力に反映されることで、偽誤情報を再生産してしまうリスクを抱えると考えられる。AI 提供者や SNS プラットフォームは、偽誤情報の流通に関与すると考えられる。学術機関や報道機関は、偽論文や偽報道が公開されることで、研究成果やニュース全体の信頼性が損なわれるという影響を受けうる。行政機関は、偽誤情報が政策立案や意思決定に混入することで、国民生活や社会制度に損害を与える恐れがある。さらに、エンドユーザーは偽誤情報の被害者であると同時に、SNS などを通じて無意識に拡散者となる点でも本影響に関連しうると想定される。

■ インタビュー調査結果

本トピックに関するインタビューは、ファクトチェックの非営利組織 I 社、報道機関 J 社に対して実施した。I 社、J 社ともに偽誤情報に対するファクトチェック業務を積極的に推進していることから、インタビューの対象組織として選定した。ここでは、インタビューから得られた要旨（回答者の発言）を記載することとし、各インタビューの詳細については A.2 のインタビュー議事を参照いただきたい。

●I 社に対するインタビュー調査において把握した事項

I 社では、インターネット等に流通する情報に対して、日々のファクトチェック活動を実施し、最新の偽誤情報の動向を社会に示す活動を実施している。また、その知見を実践的なメディアリテラシー教育、対策技術の開発企業への協力によるツール開発、実効性のある法規制の議論といった形で貢献している。

AI 由来の偽誤情報の流通状況として、I 社業務の要点ではなく、その検証も困難なため統計値等の数値はないが、実務上の感覚値として、2024 年後半から偽誤情報の質・量ともに増加傾向にあり、生成 AI の普及が影響していると考えている。また、作成される偽情報の傾向も国毎に異なっており、海外では政治関連、日本ではわいせつ物関連が多い、米国では動画の偽情報が主流だが、インドでは音声による偽情報が主流となるなど、国毎の特色が出る。加えて、偽誤情報の拡散経路は、主流のプラットフォームの変遷に伴い、X（旧 Twitter）から YouTube や TikTok へと変化してきており、災害時等の「状況の注目度」と「不確かさ」が高まる状況では、偽情報が拡散されやすい状況になることが確認されている。

偽誤情報への対応の課題として、現状技術による検証手法と限界を挙げている。具体的には、現状、偽誤情報の判別は、主に情報の描写ミスや関連情報との比較で検証されているが、現状の AI 判別ツールは検出精度が低い（例：人物の顔情報等は判別できるが、風景等の描写は判別できない等）という問題と、判別可能な範囲が狭い（ある情報が「AI によって作られたか」は判定できても、「その情報が事実か否か」という真偽の判断はできない）という問題があり、あくまで業務を補助するツールにしかなりえず、最終的には人間の目によるチェックが必要となっている。一方で偽誤情報の精度は指数関数的に進化し

ており、対策の進化よりも状況の悪化のスピードが上回っていることから、人間の目での判断が近い将来困難となるとの見立てを持っている。

上記状況を踏まえると、ファクトチェック活動だけで偽誤情報の拡散を防ぐことは今後困難となるため、ツール開発、メディアリテラシー教育、法整備など、関係ステークホルダーそれぞれが対策を強力に進める必要があるとの認識を持っている。一般市民については「画像や動画、音声があっても、それが事実とは限らない」という前提を持ち、情報に接した際は、発信源はどこか、情報の根拠は何か、関連情報はあるかの 3 点を確認する基本動作を徹底することを重要と考えており、講演・セミナー等を通して啓蒙活動を推進している。

●J 社に対するインタビュー調査において把握した事項

J 社では、報道機関として、「正確性を最優先する」という大原則の下でファクトチェック活動を実施しており、SNS や生成 AI の普及後も、報道の基本である「一次情報にあたり、裏付けを取る」という活動を徹底している。

旧来より本活動を徹底していたことから、生成 AI の普及に伴う直接的な影響は限定的な状況であり、ファクトチェック業務の負荷増大等の問題は発生していない。ただし、画像や音声のディープフェイク技術が高度化し、記者でも見抜くのが困難になっていることへの警戒感は社内で共有されており、提供された情報の確認をより慎重に行う必要があると認識している。また、社員が生成 AI を利用した結果、意図せぬ形で誤情報を記事とすることが無いよう社員が AI を利用する際は、会社が許可したサービスのみを使用し、生成物は必ずファクトチェックを行うことなどを定めている。

また、偽情報が増える中、世の中からの「真偽を検証してほしい」という要請に応えるため、報道機関の新たな役割としてファクトチェックへの取り組みを強化しており、積極的に「この情報は偽りだ」と認定し、発信する活動を実施している。特に、若い世代を中心に情報源が多様化する中、一般市民が情報の真偽を判断する能力（メディアリテラシー）を身につけることが不可欠と考えており、報道機関はその手助けをしていく役割があると認識している。本役割を具体的に実現する手段の一つとして、信頼できるメディアからの情報であることをデジタル技術で示すオリジネータープロファイルのような取り組みに参加し、読者が正しい情報源を選択しやすくする活動を行っている。また、他の報道機関やファクトチェック団体と公式・非公式に情報交換を行うことで、偽誤情報に対する感度を高めている。大学などの教育機関からの要請を受け、メディアリテラシー向上のための連携等も積極的に実施している。

●I 社、J 社のインタビュー調査結果を踏まえた考察

上記の I 社、J 社へのインタビューの結果から、インタビュー対象組織の体感では生成 AI による偽誤情報が増加しているが、生成 AI 由来の偽誤情報の統計的な情報を取得する

ことが難しいという。そのため、今後被害の実態を調査することが重要であると考えられる。また、「人間の判断」による検証手段や一般人にその判断を委ねることは限界であり、「人間の判断」＋「技術支援」のハイブリッド体制が必要であると考えられる。メディアリテラシー教育、法整備、業界標準、技術的信頼証明の確立など、各ステークホルダーが実施する対策強化を同時進行で進めるとともに、ステークホルダー間で連携を強化し対策を行っていくことが、偽情報拡散抑止と情報社会の健全化に直結すると考えられる。

3.3.2 多様性への影響

■ トピックの概要説明

生成 AI はテキスト、画像、映像などのコンテンツを大量かつ効率的に生成することを可能にし、教育や文化、ビジネスをはじめ多様な分野で活用が進んでいる。しかし、生成 AI が生み出すコンテンツが多様性を十分に反映できているかについては一部で懸念の声も見られる。学習に用いられるデータ自体が特定の属性や文化に偏っている場合、生成されるアウトプットにも同様の偏りが反映され、社会における多様性が損なわれる可能性がある。例えば、人種・性別・年齢・障害といった属性の多様性や、地域・言語・宗教など文化的多様性において偏りを含む出力が確認されている。これにより、社会的に周縁化された集団が AI 生成物から不可視化される、あるいは固定化されたステレオタイプに沿った形で表象される危険性がある。このような懸念は、生成 AI の利活用がますます進んでいくと考えられる現代社会において重要な課題であると考えられるため、本事業における検討対象としている。

さらに、このような懸念が存在する一方で、適切な設計や応用によっては、生成 AI が情報伝達やタスク遂行のバリアを軽減し、多様な人材が活躍できる場を広げるなど、ポジティブな効果をもたらす可能性も指摘されている。そこで以下では、生成 AI が多様性に対して及ぼす影響について、ネガティブな側面のみならず、ポジティブな側面についても整理している。

■ 文献調査結果

多様性については様々な学術分野で異なる定義が採用されており、定義を一意に絞り込むのは難しい。本報告書では、国際連合教育科学文化機関（UNESCO）が公表している多様性に関する資料[91]や、多様性やその関連概念の使用を分析している論稿[92]の内容に基づき、多様性を、人々や文化あるいは言語の多様な違いを認め合い尊重することとする。その上で、特に以下の 2 つの要素に分割して焦点を当てる。1 つは「属性の多様性」であり、もう 1 つは「文化の多様性」である。

第 1 に属性の多様性について、社会科学領域の独立研究者である Sadeghiani は画像生成 AI を用いた実証研究で、職業に関連する 444 枚の画像を分析した。その結果、黒人や女性、高齢者、障害者といった属性が著しく過少表象される傾向が確認された。特に科学技術分野では女性がほとんど描かれず、可視的な障害を持つ人は一度も描写されなかった。また、中高年層はステレオタイプ的に限定された役割でしか描かれなかった。こうし

[91] 文部科学省,「文化的多様性に関する世界宣言（仮訳）」
https://www.mext.go.jp/unesco/009/1386517.htm

[92] 森泉 哲,「ダイバーシティと多様性をめぐる言説の行方　―日本政府による提言のテキストマイニング分析からの考察―」
https://rci.nanzan-u.ac.jp/ninkan/publish/item/afa75b2fd332d62cfd67415df1ecb9656561471d.pdf

た結果は、生成 AI が学習データ中の偏りをそのまま再生産し、社会における属性多様性を縮減するリスクを示している [93]。

他方で、生成 AI が多様性を支援するポジティブな側面も報告されている。EY 社の国際調査によれば、障害や神経多様性を持つ従業員の 85%が「生成 AI ツールは職場をよりインクルーシブな環境にしている」と回答した。例えば会議のリアルタイム文字起こしや自動要約機能は聴覚障害者や発達障害者の作業を支援し、文章生成の補助は思考構造化が困難な人々の負担を軽減する [94]。このように適切な活用がなされれば、生成 AI は多様な人材の能力発揮を後押しし、インクルーシブな職場づくりに貢献しうる。

第 2 に文化の多様性に関しては、西洋中心の価値観を持つ AI モデルの影響が懸念されている。国際的に利用されている多くのモデルが欧米で開発されており、その結果として非西洋の文化や言語が過小に扱われる傾向がある。スタンフォード大学に所属する AI・機械学習に関する研究者の Abid らは、LLM がイスラム教徒をテロリストとして表象する傾向を指摘し、非西洋文化に対する表象的被害を示した [95]。また、コーネル大学に所属する情報科学に関する研究者の Agarwal らはインドと米国の参加者を対象にした実験で、生成 AI が提案する文章表現が米国人には有効であった一方、インド人は不自然さを感じ表現を修正する必要があったと報告している。さらにインド人は生成 AI の提案に依存することで、西洋的な文章スタイルを採用する傾向を強め、文化的均質化が進む危険があると指摘されている [96]。

次に、生成 AI による多様性への影響に関連しうるステークホルダーについて考察する。本影響に関連しうるステークホルダーとして、例えば、AI 開発者、教育機関やクリエイティブ産業、行政機関・国際機関が挙げられる。AI 開発者について、特定の属性（女性、黒人、障害者など）が十分に表象されない事例が既に確認されていることから、モデル設計において偏りを是正する責任を負う必要があると考えられる。教育機関やクリエイティブ産業は、多様性を欠いたコンテンツが広く拡散されることによって学習者や社会に偏見を再生産するリスクを抱えると考えられる。また、障害者団体やマイノリティ団体も、不適切な表象が彼らのエンパワメント活動を阻害する危険性がある。さらに、ガバナンス主体も欠かせない役割を担っている。行政機関は多様なステークホルダーを調整し、

---

[93] Sadeghiani, Ayoob, “Generative AI Carries Non-Democratic Biases and Stereotypes: Representation of Women, Black Individuals, Age Groups, and People with Disability in AI-Generated Images across Occupations.” 2025. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5343822

[94] Ernst & Young, “New research highlights benefits of Microsoft 365 Copilot for employees with disability and/or neurodivergence”
https://www.ey.com/en_uk/newsroom/2024/12/study-highlights-benefits-of-copilot

[95] Abid, Abubakar, Maheen Farooqi, and James Zou. "Persistent anti-muslim bias in large language models." 2021. https://arxiv.org/pdf/2101.05783

[96] Agarwal, Dhruv, Mor Naaman, and Aditya Vashistha. "AI suggestions homogenize writing toward western styles and diminish cultural nuances." Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025.
https://arxiv.org/pdf/2409.11360

学際的な取り組みを推進する立場にある。国際機関としての国際連合教育科学文化機関（UNESCO）は、生成 AI が特に属性に関する多様性を欠いた出力を行うことを取り扱った調査結果を公表している [97]。また、AI 倫理に関する勧告 [98]を通じ、AI ツールの設計におけるジェンダー平等確保のための具体的行動を喚起している。今後も急速に進む生成 AI の普及に対応し、国際標準やガイドラインを通じた枠組み形成が重要であると考えられる。

[97] 国際連合教育科学文化機関 (UNESCO), "Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes"
https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

[98] 国際連合教育科学文化機関 (UNESCO), "Recommendation on the Ethics of Artificial Intelligence"
https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence
e

3.4 環境への影響

3.4.1 環境への影響

■ トピックの概要説明

生成 AI は、従来のソフトウェアと比較して 1 つのタスクあたりに多くのエネルギーを消費するため、大量電力消費に伴う環境への悪影響が懸念されている。米国電力研究所(EPRI)の調査によれば、一般的な Google 検索が 1 回あたり平均 0.3Wh の電力を消費するのに対し、ChatGPT のリクエストは平均 2.9Wh を要するとされる [99]。また、スタンフォード大学の報告では、GPT-3 の学習には約 1,300MWh の電力が必要で、これは米国の 130 世帯の年間電力消費量に相当する [100]。また、より高度な GPT-4 のトレーニングには、その 50 倍もの電力が必要としたと推定されている [101]。このような状況から、生成 AI の利用が、環境に悪影響を及ぼすとの指摘があり、社会全体での重要な課題であると考えられるため、本事業における検討対象としている。

さらに、このような懸念が存在する一方で、生成 AI がエネルギー効率化や再生可能エネルギーの最適活用に貢献する事例も報告されている。そこで以下では、生成 AI が環境に与える影響について、ネガティブな側面のみならず、ポジティブな側面についても整理している。

■ 文献調査結果

生成 AI の普及に伴う環境への影響は、CO2 排出増加や電力網への負荷といったネガティブな側面と、再生可能エネルギー利用促進や効率化によるポジティブな側面の双方が確認されている。

まず、ネガティブな影響について 2 つの事例を取り上げる。第 1 に、大手テクノロジー企業の電力消費量および CO2 排出量の増加である。Google は 2030 年までに 2019 年比でデータセンター由来の CO2 排出を 50%削減する目標を掲げているが、生成 AI 利用拡大に伴い、2019 年から 2023 年の間に逆に 48%増加していると報告されている [102]。同様に Microsoft でも、データセンター拡張により 2020 年以降で電力利用に伴う CO2 排出量が

---

[99] Electric Power Research Institute, “Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption”
https://www.epri.com/research/products/000000003002028905

[100] Stanford Institute for Human-Centered Artificial Intelligence (HAI), “The AI Index 2023 Annual Report”
https://hai.stanford.edu/ai-index/2023-ai-index-report

[101] World Economic Forum, “AI and energy: Will AI help reduce emissions or increase power demand? Here's what to know”
https://www.weforum.org/stories/2024/07/generative-ai-energy-emissions/

[102] Google, “Google 2024 Environmental Report”
https://sustainability.google/reports/google-2024-environmental-report/

約 25%増加したとされる[103]。これにより、CO2 排出を伴う電力消費に関して削減目標を定め推進していたが、目標の達成が困難になってきている。第 2 に、特定地域の電力網への負荷である。アイルランドは地理的条件から欧州のデータセンター拠点となっており、Google などが大規模施設を設置している。国際エネルギー機関（IEA）のレポートによれば、同国の 2026 年の電力需要の 32%がデータセンター由来になると予測され、アイルランドのエネルギー・水規制当局は新規接続を制限するなど規制強化に踏み切っている[104]。電力需要の急増は、地域の電力供給安定性を揺るがし、住民生活や産業活動に波及する可能性がある。

一方でポジティブな影響として、AI が環境に貢献する可能性も指摘されている。MIT Technology Review によると、Google が 2024 年に公開した気象予測 AI「GenCast」は、風況を高精度に予測し、風力タービンの稼働を最適化することで発電効率を高めている[105]。また、株式会社 UPDATER と東京大学の共同研究では、AI 予測モデルを活用して発電所の翌日 24 時間の発電量を予測し、電力市場での効率的な取引を実現できると報告されている[106]。これらは AI が再生可能エネルギーの安定供給に寄与する事例である。

さらに、企業による再生可能エネルギー調達の加速も見られる。Microsoft は生成 AI による電力需要増大を背景に、ブルックフィールド・リニューアブル・パートナーズと大規模な電力購入契約を締結した。これは従来比約 8 倍の規模であり、風力・太陽光発電の普及を後押しするものと位置付けられている[107]。こうした動きは、生成 AI の利用拡大が再生可能エネルギー市場を刺激する契機となり得ることを示唆している。

次に、生成 AI による環境への影響に関連しうるステークホルダーについて考察する。本影響に関連しうるステークホルダーとして、例えば、AI 開発者、AI 提供者、AI 利用者・エンドユーザー、データセンター事業者、電力事業者が挙げられる。AI 開発者は生成 AI の運用に莫大な電力を消費するため、省エネな AI モデル開発手法や軽量化技術の確立が期待される。また、生成 AI の普及によりデータセンターの電力需要が急増するなか、Meta 社や Google 社といった大手テクノロジー企業は原子力発電への巨額投資を進めてい

---

[103] Microsoft, “Microsoft 2024 Environmental Sustainability Report”
https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report/

[104] International Energy Agency, “Electricity 2024 – Analysis and forecast to 2026”
https://www.iea.org/reports/electricity-2024

[105] MIT Technology Review, “Google DeepMind’s new AI model is the best yet at weather forecasting”
https://www.technologyreview.com/2024/12/04/1107892/google-deepminds-new-ai-model-is-the-best-yet-at-weather-forecasting/

[106] 東京大学,「みんな電力×東京大学、AI 予測モデルを用いた発電量予測システムの予測精度向上に取り組む～FIP 導入に伴い、アグリゲーターとしての運用も開始～」
https://www.t.u-tokyo.ac.jp/press/pr2022-03-24-001

[107] Brookfield Renewable Partners, “Brookfield and Microsoft Collaborating to Deliver Over 10.5 GW of New Renewable Power Capacity Globally”
https://www.globenewswire.com/news-release/2024/05/01/2873042/0/en/Brookfield-and-Microsoft-Collaborating-to-Deliver-Over-10-5-GW-of-New-Renewable-Power-Capacity-Globally.html

るとの報道もある[108]。AI 提供者は、生成 AI の利用規模やシステム設計の在り方が電力消費量を大きく左右すると想定されるため関連しうる。AI 利用者・エンドユーザーは、利用が増えるほど電力消費は拡大するが、個々人が自らの利用が環境に与える影響を把握するのは難しく、主に利用量を通じて間接的に関与するに留まると想定される。データセンター事業者は、インフラ提供者として電力消費や冷却効率に直結しており、立地や建物構造、設備設計によって地域の環境負荷が大きく変わるため関連しうる。電力事業者は、AI 関連需要の増大により新たな市場機会を得ており、AI 開発者やデータセンター事業者との連携を通じて、再生可能エネルギーの導入や供給体制の拡充を進める役割を担っていると想定される。

[108] ハーバードビジネスレビュー,「原子力発電に投資する巨大テック企業の思惑」
https://dhbr.diamond.jp/articles/-/12229

4　今後の検討に向けて

本章では、今後の検討に向けて、本調査で得られた結果を踏まえて考えられる AI セーフティに関する社会技術的影響に関連する現状の課題と対応に関する考察を記載する。

4.1 AI セーフティに関する社会技術的影響に関連する現状の課題

本調査で得られた結果を踏まえ、AI セーフティに関する社会技術的影響に関する現状の課題と考えられる事項を記載する。生成 AI の急速な普及は、犯罪への悪用、法的課題、さらには経済格差といった社会構造の変化に至るまで、多岐にわたる問題を顕在化させうる。ここでは、上述したような課題を、技術的課題、社会的課題、制度的課題に分けて記載することとする。なお、本報告書で取り上げる論点は、調査結果から抽出された代表的課題に基づくものであり、すべての要因を網羅的に列挙するものではない点に留意する必要がある。

■ 技術的課題

技術的な課題について、ここでは生成 AI の学習データの品質、不適切な出力の多様な影響の 2 つに言及する。まず、学習データの品質について記載する。生成 AI はインターネット上の文書等、膨大なデータを無作為に学習に利用しているモデルも存在する。生成 AI の出力結果は学習データに大きく依存するため、学習データの品質が出力の品質に大きく関連すると考えられる。例えば、わいせつ物の生成や流通（3.1.2 節）のインタビュー調査結果から、エンドユーザーが特定の人物の画像を入力や学習に利用しない、一からプロンプトにより出力させるもので、意図せず特定の人物に近いわいせつ画像が出力されてしまう事例が国内外で発生しているとされている。これは、AI 開発者がデータを無作為にモデルの学習に利用しているために、特定個人の画像が本人の同意なしに学習に利用されていることを示唆していると考えられる。また、特に児童を対象とするわいせつ物は児童性的虐待として国際的な問題となっており、同様にわいせつ物の生成や流通（3.1.4 節）のインタビュー調査結果において、わいせつ物を含まないデータによる学習を期待する声もある。

さらに、出力のバイアスが社会に及ぼす影響（3.1.2 節）の文献調査結果に示した通り、マルチモーダルデータセットの規模拡大に伴い、生成 AI の出力によって人種・民族といった社会的要素に関する差別が生じる可能性が高まると指摘する文献も存在する。また、AI 事業者ガイドライン（第 1.1 版）でもバイアスを生み出す要因の 1 つとして学習データを挙げており、AI 開発者はデータの質を管理するための相当の措置を講じることが重要とされている [109]。このように、AI 開発者がモデルの学習にどのようなデータを利用するか

[109] AI 事業者ガイドライン（第 1.1 版）
https://www.soumu.go.jp/main_content/001002576.pdf

は生成 AI が社会に与える影響に大きく関連するため、重要な課題であると考えられる。

次に、不適切な出力の多様な影響について記載する。生成 AI はその汎用性から、自然言語生成・翻訳・プログラミング・画像生成など多様なタスクに対応できる。これは利便性を高め利用用途が増える一方で、不適切な出力が多様な影響を及ぼす可能性もあると考えられる。例えば、サイバー攻撃への悪用（3.1.3 節）や偽誤情報の生成や拡散（3.3.1 節）の文献調査結果やインタビュー調査結果に示したように、生成 AI が悪用されることで、フィッシングメール文、偽記事や悪意あるディープフェイクといった不適切な出力がなされる懸念がある。さらに、生成 AI への過信が及ぼす心身への影響（3.1.1 節）、機密情報の学習や漏洩の懸念（3.2.5 節）の文献調査結果から、エンドユーザーに不適切な誘導を行い心身に影響をもたらした事例や AI の出力を通じて機密情報が漏洩した事例も存在する。このような事例から、不適切な出力が多様な影響を及ぼす可能性があり、その出力制御が課題であることが示唆される。そのため、イノベーションを促進させつつ、AI 開発者はこのような AI の悪用、不適切な誘導、機密情報の漏洩といった事象が発生しないよう出力を制御することが重要であると考えられる。

■ 社会的課題

社会的な課題について、ここでは格差の拡大、個人のリテラシーの不足、という 2 つの課題に言及する。まず、格差の拡大について記載する。経済格差へ及ぼす影響（3.2.4 節）の文献調査結果から、生成 AI が社会に普及しているものの、活用が進んでいる個人・企業・国家とそうでない主体が存在すると考えられる。今後生成 AI の利活用が社会としてますます進んでいくことが想定されるため、生成 AI の活用格差が大きな経済格差へと波及する懸念があると考えられる。さらに、労働環境への影響（3.2.2 節）の文献調査結果から AI が人間に取って代わって仕事をするとして、人員削減を計画している事業者も存在する。そのため、AI に代替される職業に従事している者とそうでない者の間で格差が拡大する懸念がある。

次に、個人の AI リテラシー不足について記載する。生成 AI への過信が及ぼす心身への影響（3.1.1 節）や偽誤情報の生成や拡散（3.3.1 節）の文献調査結果やインタビュー調査結果から、エンドユーザーが生成 AI の特性や限界を十分に理解しないまま利用することで、生成 AI への過信や生成 AI の誤用が進行する危険性があると考えられる。また、生成 AI により容易に偽情報の生成が可能となっていることから、生成 AI を利用していない個人を含め、偽情報のより早く広い範囲への拡散に影響しうる。さらに、偽誤情報の生成や拡散（3.3.1 節）のインタビュー調査結果から、人々は偽誤情報の内、14.5%しか誤っていると認識できないという研究結果があると言及する事業者も存在する。

加えて、同様に偽誤情報の生成や拡散（3.3.1 節）の文献調査結果やインタビュー調査結果から、SNS の普及によって個人のリテラシー不足による AI が社会に与えるネガティブな影響を増大させうることも示唆される。近年 SNS が普及し、例えば災害等の社会的イン

パクトの大きなトピックに関する情報は容易に拡散するようになっている。このような社会的要素に加え、生成 AI はもっともらしい画像や動画を容易に生成できることと、SNS 利用者の AI リテラシーの不足が相まって、容易に偽誤情報が拡散しうる。さらに、同様に偽誤情報の生成や拡散（3.3.1 節）の文献調査結果からわかるように、SNS や検索エンジンのアルゴリズムは SNS 利用者の関心を優先し、エコーチェンバー現象を助長しうることも指摘されている。このように SNS の普及により生成 AI が社会に与える影響は大きくなっていると考えられる。以上より、個人の AI リテラシー不足は重要な課題であると考えられる。したがって、個人の AI リテラシーを高めることで、これらの影響は軽減でき、生成 AI をより安全かつ有益に活用できる可能性が広がると期待される。

■ 制度的課題

本調査では、生成 AI と著作権の関係や、わいせつ物に関する制度的課題を指摘する文献やインタビュー結果が得られた。

著作権に関しては、生成物の権利に関する懸念（3.2.1 節）の文献調査結果から、海外では生成 AI により作成された画像に対して著作権侵害が認められた事例が存在している。また、生成 AI による作風の模倣を著作権の保護対象とすべきか否かの議論が国内外で発生した事例も存在している。文化庁によれば、著作権制度は生成 AI と著作権の関係に関する判例および裁判例の蓄積が少なく、判断が容易でないと考えており、これを受けて AI と著作権に関する考え方を記す文書を公表している [110]。なお、当該文書でも言及されている通り、本文書は生成 AI と著作権の関係についての一定の考え方を示すものであって、本考え方自体が法的な拘束力を有するものではない。当該文書では「AI をはじめとする新たな技術への対応については、著作権法の基本原理や、法第 30 条の 4 をはじめとする各規定の立法趣旨といった観点からの総論的な課題を含め、中長期的に議論を行っていくことが必要と考えられる」としており、生成 AI による生成物の著作権に関する問題については引き続き検討を行っていく必要があると考えられる。

また、わいせつ物の生成や流通（3.1.4 節）のインタビュー調査結果から、児童ポルノ禁止法は「実在児童」のみを規制対象とするため、「非実在児童」の生成物を取り締まりにくい状態になっていることが指摘されている。さらに、生成 AI 特有の「一部実在児童（顔や体だけ実在児童のディープフェイク）」の取り締まりが容易ではないことも指摘されている。これは、顔や体といった児童の一部が利用されていると考えられる場合であっても、実在性の解釈が多様であり、実在性の判断が困難であることに起因しているとされている。また、同インタビュー調査結果からは、相談者が法執行機関に相談に行っても対応が難しいと告げられた事例も存在するとされている。そのため、AI 時代の制度設計の継続議論が必要であると考えられる。

[110] 文化庁,「AI と著作権に関する考え方について」
https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/pdf/94037901_01.pdf

これまで述べた AI セーフティに関する社会技術的影響に関連する現状の課題例と分類の関係を表 4 に示す。

表 4 AI セーフティに関する社会技術的影響に関連する現状の課題例

#	分類	課題
1	技術的課題	学習データの品質
2		不適切な出力の多様な影響
3	社会的課題	格差の拡大
4		AI リテラシー不足
5	制度的課題	AI 生成物の取扱い

4.2 AI セーフティに関する社会技術的な影響への対応に関する考察

4.1 節において記載した現状の課題認識を踏まえ、AI セーフティに関する社会技術的な影響について、今後実施を検討することが望ましいと考えられる対応を記載する。4.1 節で整理した課題を踏まえ、AI セーフティの社会技術的影響に対する対応は、国・地域レイヤ、企業レイヤ、個人レイヤで対応を進めることが求められると想定される。なお、本報告書で取り上げる対応は、調査結果から抽出された事項の考察に基づいたものであり、すべての対応を網羅的に列挙するものではない点に留意する必要がある。

■ 国・地域レイヤによる対応

国・地域レイヤによる対応として、表 4 で言及した課題#3（格差の拡大）、#4（AI リテラシー不足）、#5（AI 生成物の取扱い）への対応を考察する。格差の拡大、AI リテラシー不足、制度観点の AI 生成物の取扱いへの対応は個人や企業レイヤだけでの対応では限界があり、国・地域レイヤで対応する必要があると考えられるため取り上げる。

まず、格差の拡大については、生成 AI への過信が及ぼす心身への影響（3.1.1 節）のインタビュー調査結果から、初等教育や中等教育時等、できるだけ早く AI に触れることができるようにすることを指摘する事業者が存在する。これは、当該インタビュー調査結果からも示唆されるように、幼少期から AI に触れてきた人と触れてこなかった人とで AI の使い方の習熟度合に差が開き、結果として格差に繋がる可能性があるためとされている。そのため、今後 AI の利用の活用度合いが格差にどのような影響を及ぼすかの効果検証を行う等が対応として考えられる。

次に、AI リテラシー不足への対応については、上述した格差の拡大への対応にも関連しており、教育現場への AI の活用を検討し、AI リテラシーを向上させていくことが 1 つの対応として考えられる。さらに、教育現場での対応のみならず、社会全体の AI リテラシ

ーを向上させるために、本事業のような調査研究結果の啓蒙や、AI を基軸とした働き方の指針を提示していくことも考えられる。なお、内閣府が公開する「人工知能基本計画骨子（たたき台）」では、AI 社会に向けた継続的変革の具体的な取組例として、初等中等教育や一般市民における AI リテラシー向上支援や AI 時代の働き方の検討を挙げている[111]。このような多角的なアプローチを通じて社会全体の AI リテラシーを向上させることで、AI が社会に与えるネガティブな影響を低減させることができうると考えられる。

最後に、AI 生成物の取扱いについては、AI が今後ますます普及することを前提とし、本事業で取り上げた様々な影響を考慮した制度設計の議論を継続的に行っていくことが考えられる。その際、本事業でインタビュー調査に協力いただいたような現状の実態を把握した有識者を巻き込み、わいせつ物の生成や流通（3.1.4 節）のインタビュー調査で指摘されているような運用面も考慮した制度設計を行うことが考えられる。なお、2025 年 9 月、政府はインターネットの利用を巡る青少年の保護の在り方に関する関係府省庁連絡会議において、生成 AI により生成されたわいせつ物を含む対応に関する工程表を公開しており、今後対策の実効性を高めるための方策の在方が検討されるとされている[112]。また、前提として、日本の人工知能関連技術の研究開発及び活用の推進に関する法律（AI 法）にもある通り、イノベーションを促進しつつ AI が社会に与えるネガティブな影響に対応することが重要であると考えられる[113]。

■ 企業レイヤによる対応

まず、AI 開発者、提供者の観点で課題#1（学習データの品質）、課題#2（不適切な出力の多様な影響）について記載する。学習データの品質においては、例えば生成物の権利に関する懸念（3.2.1 節）や出力のバイアスが社会に及ぼす影響（3.1.2 節）、多様性への影響（3.3.2 節）の文献調査結果からも示唆されるように、著作者が許可を与えたデータのみを利用したり、偏見を低減させたり、多様性に富んだデータの拡充を進めたりすることが重要であると考えられる。一方、生成 AI の学習には大量のデータが必要になり、データ 1 件 1 件を精査することは難しいと想定されるため、どのように学習データの品質を高めていくかは今後継続議論が必要であると考えられる。

不適切な出力の多様な影響に関する対応について、例えば継続的にガードレールの調整を行うことが考えられる。既に大手テクノロジー企業はガードレールの強化に取り組んで

---

111 内閣府，「人工知能基本計画骨子（たたき台）」
https://www8.cao.go.jp/cstp/ai/ai_hq/1kai/shiryo2_2.pdf

112 インターネットの利用を巡る青少年の保護の在り方に関する関係府省庁連絡会議，「「課題と論点の整理」に基づく工程」
https://www.cfa.go.jp/assets/contents/node/basic_page/field_ref_resources/b6706386-18be-48af-adb6-0813bdbbd0fe/983a093e/20250926_councils_internet-kaigi_b6706386_01.pdf

113 内閣府，「人工知能関連技術の研究開発及び活用の推進に関する法律（Ａ Ｉ 法）」
https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html

いる[114]が、AI 技術の進展や世の中の流れの変化に伴いリスクも変化すると想定されるため、ガードレール機能を調整していくことで、AI が社会に与えるネガティブな影響を低減させることができると考えられる。さらに、サイバー攻撃への悪用（3.1.3 節）のインタビュー調査結果から示唆されるように、AI の生成物にコンテンツの信頼性を証明する技術を用いて来歴情報を付与することで AI による生成物であることを可視化することも対応として考えられる。このような対応を行うことで、不適切な出力を未然に防いだり、一般市民が AI による生成物であると気づきを与えたりすることができ、生成 AI によるサイバー攻撃や偽誤情報の拡散等の多様な影響を抑制することができると考えられる。

次に AI 利用者の観点で課題#3（格差の拡大）、#4（AI リテラシーの不足）について記載する。格差の拡大については、「■ 国・地域レイヤによる対応」で言及した内閣府が公開する「人工知能基本計画骨子（たたき台）」[115]でも言及されているように、AI 社会から取り残されないよう人間力を向上させることが重要であると考えられる。そのために、企業内で AI 時代の働き方を検討し、積極的に AI の利活用を推進していくことが重要であると考えられる。一方 AI の利活用にはリスクも存在するため、労働環境への影響（3.2.2 節）のインタビュー調査結果から得られたように、組織内で AI 倫理ポリシーを定め、適切にリスク管理することが重要であると考えられる。その上で、同様に労働環境への影響（3.2.2 節）の文献調査結果から示唆されるように、例えば職務をタスク単位に分解し、「代替可能な領域」と「補完が有効な領域」とを区別することによって、リスキリングの重点を戦略的に特定し、組織固有の学習データを活用することで、持続的な競争力の確保につなげることが考えられる。上記を達成するためにも、組織配下のエンドユーザーの AI のリテラシーを向上させることも欠かせないと考えられる。組織全体で AI 利活用を促進するために、組織内で AI に関する教育や AI の利活用推進を通じて AI リテラシーを向上させ、組織全体の AI リテラシーを向上させることで他組織との格差の拡大を緩和させることができると考えられる。

■ 個人レイヤによる対応

個人レイヤによる対応として、課題#3（格差の拡大）、#4（AI リテラシーの不足）への対応を記載する。個人レイヤでも AI に関するリテラシーを向上させるために研鑽を積むことが重要であると考えられる。例えば、偽誤情報の生成や拡散（3.3.1 節）のインタビュー調査結果から示唆されるように、生成 AI が必ずしも事実に基づく出力を行うわけではない等、AI の能力と限界を正しく理解することが考えられる。また、同様に偽誤情報の生

[114] AWS,「Amazon Bedrock のガードレールが、新しい機能により、生成 AI アプリケーションの安全性を強化」
https://aws.amazon.com/jp/blogs/news/amazon-bedrock-guardrails-enhances-generative-ai-application-safety-with-new-capabilities/

[115] 内閣府,「人工知能基本計画骨子（たたき台）」
https://www8.cao.go.jp/cstp/ai/ai_hq/1kai/shiryo2_2.pdf

成や拡散（3.3.1 節）のインタビュー調査結果から、情報に接した際は、発信源はどこか、情報の根拠は何か、関連情報はあるかの 3 点を確認することが重要であるという見解が得られている。さらに、AI の進化や社会の変化に対応するため、主体的に学び続ける姿勢を持ち、自身の専門分野に AI をどう活用できるかを考え、新たなスキルを習得することが必要であると考えられる。結果として、個人レイヤでの格差の緩和にもつながると考えられる。国・地域レイヤ、企業レイヤでの対応も重要であるが、個人一人一人による対応を行わないと効果を最大化することは難しいため、個人レイヤの対応も必要であると考えられる。

上述の通り、AI セーフティに関する社会技術的影響に関連する課題への対応は、国・地域、企業、個人といった様々なレイヤで行うことが必要であると考えられる。繰り返しとなるが、AI セーフティに関する社会技術的な影響は、AI に関する技術的な環境の変化や、それを取り巻く社会的な環境の変化の影響を受けて急速に変化し得る。そのため、AI セーフティに関する社会技術的な影響やその課題、対応については継続的な検討を行うことが重要である。

A　付録

A.1　概要調査対象一覧

以下に文献調査において概要調査を行った文献の一覧を示す。なお、「3 調査結果」で記載した通り、報道は「倫理・法」、「経済活動」、「情報空間」、「環境」のうち、どの影響に主に関連しうるかで分類しているが、学術論文、レポートは 1 文献内に複数の事例について言及されているため、分類していない。

■　報道

倫理・法への影響に関連

- CNN,「会計担当が３８億円を詐欺グループに送金、ビデオ会議のＣＦＯは偽物　香港」
  https://www.cnn.co.jp/world/35214839.html
- NHK,「生成 AI と会話を続けた夫は帰らぬ人に」
  https://www3.nhk.or.jp/news/html/20230728/k10014145661000.html
- 集英社オンライン,「＜全国初の摘発＞AI 生成“わいせつ”画像を 9000 点以上出品、“素人”男女 4 人が逮捕「原価が安く、稼げた」深刻化するディープフェイク」
  https://shueisha.online/articles/-/253693
- ABEMA TIMES,「自称“AI 誘発性心理反応”の 30 代ニート「誰にも肯定されていない不安感に」 生成 AI の“誤った使い方”に専門家が警鐘」
  https://times.abema.tv/articles/-/10179735?page=1
- 日刊 SPA!,「“人間よりも AI”にお悩み相談する人が増加中。専門家が明かす、AI に頼りすぎる危険性」
  https://nikkan-spa.jp/2093701
- 読売新聞,「生成ＡＩ悪用しウイルス作成、警視庁が２５歳の男を容疑で逮捕…設計情報を回答させたか」
  https://www.yomiuri.co.jp/news/national/20240528-OYT1T50015/
- Yahoo! JAPAN,「「AI とのチャットに依存、14 歳が死亡」母親が提供元を提訴、その課題とは？」
  https://news.yahoo.co.jp/expert/articles/7225ddf3ec2e66fae6a09bd6cc96313b2a44e6f8
- 日本経済新聞,「生成 AI 悪用、楽天回線 1000 件不正契約か　中高生を逮捕」
  https://www.nikkei.com/article/DGXZQOUE271BZ0X20C25A2000000/?msockid=226434b3851266c3346c217884e067dc
- 読売新聞,「事件・事故の犠牲者の顔写真、生成ＡＩが無断使用…遺族「使うのやめて」・識者「尊厳にかかわる」」

https://www.yomiuri.co.jp/national/20240407-OYT1T50068/

- 読売新聞,「宿題もリポートも生成ＡＩが作った「正解」丸写し、教諭は嘆く「これじゃ無料の代行業者だ」」
  https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20240430-OYT1T50001/
- 共同通信,「AI 悪用か、社長の偽音声で指示　部下に電話、不正送金命じる」
  https://www.47news.jp/12325929.html
- 読売新聞,「中学１年生２５０人の半数超、理科の課題で同じ間違い…教諭の違和感の正体は生成ＡＩの「誤答」」
  https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20240306-OYT1T50080/
- 東洋経済新聞社,「香港警察が摘発「ディープフェイク詐欺団」の手口」
  https://toyokeizai.net/articles/-/835536
- YTV NEWS NNN,「【賛否】癒しか冒涜か―AI で亡き人を再現『AI 故人』サービス相次ぎ登場　“メッセージ型”に“対話型”　願い叶える新たな弔い方」
  https://news.ntv.co.jp/n/ytv/category/society/yt8f7d5827d0564c8b8296c12885bf14cc
- ITmedia,「生成 AI で福岡の PR 記事作成→“架空の祭りや景色”への指摘が続出　開始 1 週間で全て削除する事態に」
  https://www.itmedia.co.jp/aiplus/articles/2411/08/news167.html
- 日テレ NEWS NNN,「生成 AI でニュースを偽造？　日テレ番組悪用の詐欺広告　一体どうやって？　#みんなのギモン」
  https://news.ntv.co.jp/category/society/79f52d1bd558460ca350b160ae7c7b50
- セキュリティ対策 Lab,「ユーロポール、AI を悪用した児童虐待コンテンツを摘発、19 カ国の国際捜査で 25 人逮捕」
  https://rocket-boys.co.jp/security-measures-lab/europol-ai-generated-child-abuse-crackdown-25-arrested/
- New Straits Times “University students expelled for failing to disclose AI use“
  https://www.nst.com.my/world/world/2025/03/1192401/university-students-expelled-failing-disclose-ai-use

経済活動への影響に関連

- 東洋経済新聞社,「「これは営業じゃない! 業務妨害だ」 迷惑な"AI 営業"に中小企業の社長が激怒したワケ《背景に AI の利用拡大》送信先企業にもたらされる 3 つの問題」
  https://toyokeizai.net/articles/-/870844?page=3
- TBS NEWS DIG,「チャット GPT の新機能で生成した「ジブリ風」画像が SNS で流行　著作権侵害の懸念も」
  https://newsdig.tbs.co.jp/articles/-/1817549?display=1

- 共同通信,「声優の AI 音声、無断利用に警鐘　経産省、違反の恐れを例示」
  https://www.47news.jp/12559948.html
- 千葉日報オンライン,「新聞協会、ＡＩは「著作権侵害」　検索連動型、記事の利用承諾要請」
  https://www.chibanippo.co.jp/newspack/20240717/1250239
- 共同通信,「「AI エージェント」に注目　自ら判断、仕事を代行」
  https://www.47news.jp/12652423.html
- 読売新聞オンライン,「生成ＡＩで日本人の研究者かたり論文捏造か、収入目的の海外サイト「ハゲタカジャーナル」に掲載」
  https://www.yomiuri.co.jp/national/20241120-OYT1T50136/
- 神奈川新聞,「ＡＩで「エヴァ」のポスター生成し販売　神奈川初、容疑で男性２人書類送検」
  https://www.kanaloco.jp/news/social/case/article-1142687.html
- MONOist,「60%が生成 AI を業務で利用、そのうち 85%が「人に頼らず AI でいいや」」
  https://monoist.itmedia.co.jp/mn/articles/2501/30/news096.html
- 産経ニュース,「デジタル赤字６・６兆円に拡大、止まらぬ国富流出も生成 AI 追い打ちに　問われる成長戦略」
  https://www.sankei.com/article/20250210-HZVXV6AXORK3VAQ35BILO5VEJU/

情報空間への影響に関連

- ITmedia,「自筆した論文が勝手に解説動画にされた？→実は存在しない“フェイク論文”　著者名を無断利用、生成 AI を悪用か」
  https://www.itmedia.co.jp/aiplus/articles/2504/09/news059.html
- TechnoEdge,「新たな AI 音声生成ツール、公開直後から著名人の声でヘイトスピーチや不適切発言させるディープフェイクボイスが横行」
  https://www.techno-edge.net/article/2023/02/01/795.html
- Gigazine,「ウソの通報で特殊部隊を送りこむ嫌がらせ「スワッティング」を完全自動化した代行業者が登場している」
  https://gigazine.net/news/20230414-torswats-swatting-automated/
- 読売新聞,「生成ＡＩで関東大震災「新証言」を作成…「捏造」批判受け日赤の企画展中止」
  https://www.yomiuri.co.jp/national/20230903-OYT1T50216/
- 共同通信,「「マジで悲惨すぎる…」被災の画像、実はディープフェイクだった　高まる生成ＡＩの悪用懸念にどう向き合う？」
  https://www.47news.jp/relation-n/2024103006

- 読売新聞オンライン,「ＮＨＫがネット配信ニュースでＡＩ翻訳ミス…「尖閣諸島」を「釣魚島」と表示、多言語字幕サービス終了」 https://www.yomiuri.co.jp/culture/tv/20250212-OYT1T50154/
- 読売新聞,「生成ＡＩで岸田首相の偽動画、ＳＮＳで拡散…ロゴを悪用された日テレ「到底許すことはできない」」 https://www.yomiuri.co.jp/national/20231103-OYT1T50260/
- 読売新聞,「生成ＡＩの偽動画、偏った考え増幅の「エコーチェンバー」引き起こす…安倍元首相の追悼行事で騒動引き起こす」 https://www.yomiuri.co.jp/national/20231207-OYT1T50052/
- Reuters,「焦点：米で人気の中国発ニュースアプリ、ＡＩ生成で誤報と「創作」連発」 https://jp.reuters.com/markets/japan/J6FZL7YC35MLZPUHXJIZF7OZPU-2024-06-06/
- ITmedia,「AI が「不適切な内容」公式 X に投稿、自社掲示板の内容を要約　マンション情報サイト運営元が謝罪」 https://www.itmedia.co.jp/aiplus/articles/2506/02/news073.html
- Yahoo! JAPAN,「「AI が生成したゴミでネットが汚染された」研究用データベースが更新停止したわけとは？」 https://news.yahoo.co.jp/expert/articles/b8099311e535ba29b1b35dc47a74ee7c5b00ad0e
- Gadget Gate,「NY 市の AI チャットボット「MyCity Chatbot」、「危険なほど不正確」な情報提供で批判を浴びる」 https://gadget.phileweb.com/post-72785/
- GIZMODO,「チョコレートの夢はどこ？お粗末すぎた『ウォンカ』の没入型イベント」 https://www.gizmodo.jp/2024/03/wonkas-ai-immersive-event.html
- JBpress,「生成 AI に騙される弁護士がいまだに相次ぐ――裁判に架空の判例を提出した弁護士には制裁金の勧告」 https://jbpress.ismedia.jp/articles/-/86872
- Gigazine,「「チャットボットの誤回答に責任はない」と弁解していたエア・カナダに裁判所が損害賠償支払いを命令」 https://gigazine.net/news/20240219-air-canada-chatbot-mistake/

環境への影響に関連

- 日本経済新聞,「JERA、データセンター向けガス火力発電へ　AI 過熱受け」 https://www.nikkei.com/article/DGXZQOUC2260L0S5A420C2000000/?msockid=27

b82f227b2e6ebb136b3a9b7a546fe8
- MIT News, "Explained: Generative AI's environmental impact" https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117

■ **学術論文**

- Smith, Jessie J., et al. "The Generative AI Ethics Playbook." arXiv preprint arXiv:2501.10383 (2024). https://arxiv.org/abs/2501.10383
- Marchal, Nahema, et al. "Generative AI misuse: A taxonomy of tactics and insights from real-world data." arXiv preprint arXiv:2406.13843 (2024). https://arxiv.org/abs/2406.13843
- Weidinger, Laura, et al. "Sociotechnical safety evaluation of generative ai systems." arXiv preprint arXiv:2310.11986 (2023). https://arxiv.org/abs/2310.11986
- Weidinger, Laura, et al. "Star: Sociotechnical approach to red teaming language models." arXiv preprint arXiv:2406.11757 (2024). https://arxiv.org/abs/2406.11757
- Fazelpour, Sina, and Maria De-Arteaga. "Diversity in sociotechnical machine learning systems." Big Data & Society 9.1 (2022): 20539517221082027. https://journals.sagepub.com/doi/full/10.1177/20539517221082027
- Bastani, Hamsa, et al. "Generative ai can harm learning." The Wharton School Research Paper (2024). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4895486
- 美馬のゆり. "AI の社会的影響と教育の転換." 名古屋高等教育研究 25 (2025): 11-24. https://web.cshe.nagoya-u.ac.jp/publication/journal/img/no25/02.pdf

■ **レポート**

- 研究開発戦略センター,「人工知能研究の新潮流 2025～基盤モデル・生成 AI のインパクトと課題～」 https://www.jst.go.jp/crds/pdf/2024/RR/CRDS-FY2024-RR-07.pdf
- 株式会社大和総研,「生成 AI が日本の労働市場に与える影響①」 https://www.dir.co.jp/report/research/economics/japan/20231208_024132.pdf
- 株式会社情報通信総合研究所,「生成 AI 時代におけるフェイクニュースとの向き合い方」 https://www.icr.co.jp/newsletter/wtr430-20250130-eshimizu.html
- 国立研究開発法人 新エネルギー・産業技術総合開発機構（NEDO）, みずほリサーチ

&テクノロジーズ株式会社,「生成 AI の著作権侵害等のリスクとその低減技術動向調査結果」
https://www.nedo.go.jp/content/100977000.pdf

- 日本学術会議,「生成 AI を受容・活用する社会の実現に向けて」
https://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-26-t381.pdf

A.2　インタビュー調査

本節では、インタビュー調査対象組織（業種）の一覧と対象組織から許諾が得られたものについて、インタビュー議事を記載する。調査対象トピックとインタビュー対象組織（業種）の一覧を表 5 に示す。

表 5 調査対象トピックとインタビュー組織（業種）一覧

通番	調査トピック	インタビュー組織（業種）
A	生成 AI への過信が及ぼす心身への影響	教育事業者の社内シンクタンク
B		大学
C	サイバー攻撃への悪用	情報セキュリティベンダー
D		情報セキュリティベンダー
E	わいせつ物の生成や流通	児童保護団体
F		ネットパトロール団体
G	労働環境への影響	金融サービス業
H		IT サービス業
I	偽誤情報の生成や拡散	ファクトチェックを行う非営利組織
J		報道機関

以下に公開が許諾されたインタビュー議事を記載する。

A 社（教育事業者の社内シンクタンク）

① 生成 AI の利用実態と学習方法の変化

・ 学校での AI の利用（生徒による利用）は小・中・高・大のそれぞれでどの程度進んでいるか。何らか情報をお持ちであればご教示いただきたい。（インタビュアー）

→ 前提として、AI は変化が早いため、定量的な調査は行っていない。あくまでも定性的な感触となる。ここで言う AI の利用は、学校のみでの利用に限るのか、それとも家庭での利用も含むのか確認したい。（A 社）

→ 一番聞きたいところは、学習でどの程度利用されているかとなるが、学校での利用と家庭での利用で異なるところがあればお聞きしたい。（インタビュアー）

→ 承知した。学齢が上がるにつれて、使用頻度は高くなると認識している。就職する大学生はほとんどが使用しており、企業に出すエントリーシート作成に AI を活用している。また、調査のほか文章構成、文章の推敲などに使用している。レポート課題での活用が多いのではないか。大学生は、日常的に AI をツールとして利用している

が、大学 1 年生の段階ではまだ使用頻度は低いように思われる。一方で、大学の授業の中での AI 利用を明確に禁止している大学もある。そのほか、専門学校、芸術系大学で利用している事例はある。高校生は、何割使用しているかは不明であるが、かなり使っている印象を持っている。プログラミング教育などだけの使用ではなく、インターネット検索ではサーチエンジンを検索窓口として使用せず、ChatGPT などの AI を検索窓口として使用している。高校では、先生の判断で AI を使用して様々な取り組みが行われており、情報の授業の時間などで触れたりしている。コンピュータグラフィック、動画や文書作成などの取り組みが行われていると認識している。一方で、ChatGPT を使用して英会話のトレーニングをするなど作りこまれた AI のアプリケーションを使うことはあるが、学生がプロンプトを工夫して AI を高度に使用している事例はあまり聞いていない。（A 社）

→ 小学校について、多くの学校ではまだ教育に AI が利用されていない。中学校は、保護者に許可取れれば使用が可能で、要約・英作文の推敲・調べ学習に使用している学校もある。学校での AI の使用について、付加価値が高い良いユースケースはまだあまり出てきていない。今は AI を体験するために利用することが多い状況と考える。（A 社）

・ 小・中・高・大の学校において、どのような場面で生成 AI を活用（生徒による利用）されているか。もしくは、どのような場面で活用することが効果的であると考えているか。（インタビュアー）

→ 生徒は、情報の整理分析、編集制作、文字の構成を変えるなどに AI を多く利用している。現状学校の授業で利用する場合は、AI の概要を理解したり、AI と簡単なやり取りをしたりイラストをつくるなどの初歩的なケースが多い。今後は教員が生徒に AI を利用し自学自習させることで自身の時間を作り、指導が必要な生徒の個別対応に回ることが考えられる。AI を利用して自学自習した生徒は、利用記録が残っているため、学習進度の理解の裏付けとすることができる。あるいは、利用記録（ログ）を AI で分析し、生徒の理解度を判断が可能となる。これはある程度作り込まれた AI を使わなければできないことだが、教員不足や児童・生徒の多様化が進む中で期待されている使い方である。（A 社）

→ 小学生の活用例として、現在は授業より校務（企画書、保護者への案内文など）で使用している。中学校では英語翻訳などでも利用が見られる。高校では、要約の比較、英作文の推敲、探究テーマの磨き込み、プログラミングのコード作成、面接・小論の模擬相手、相談相手などでの利用が考えられる。（A 社）

・ 上記のような生成 AI の活用によって、学習の方法に変化は発生しているか。（インタビュアー）

→ 良いところは、アイスブレークや注目を集めるなどのアクセントとなる。エンターテインメントの要素としては、成功している。また、生徒が使用する場合、長い文章を書くことが苦にならなくなった。その結果、自身で長い文章を考える力が伸びていかない懸念はあるが、利用するメリットが多ければ、問題ない見方もある。(A 社)

→ 人と話をするより、AI と話をする方が良いと言う中高生が現れている。2024 年の話だが、話の内容によって複数の AI を使い分けていると言う話を聞いたことがある。学習の面では、英作文を AI に任せても、自ら英作文を学び続ける生徒は稀だと考える。これからは、より意欲や自分で継続できる動機付けによって差がつくことが考えられる。(A 社)

→ 例えば、算数の例題をもとに、問題を AI が生成するところが強みと考えていたが、プライベートでこのような使い方はあまり意識されていない理解で良いか。(インタビュアー)

→ 問題作成というより、インタラクティブな学習の感覚である。(A 社)

・ 宿題等で生成 AI を利用しない前提の場面もあると考えているが、自宅等学生が見えない場面での生成 AI 利用の統制をどのように図ることが考えられるか。(インタビュアー)

→ 考え方の問題である。AI を利用した上でのパフォーマンスを見ていくべきであると考えている。大学では AI を使用することがベースとなり、中長期的に見た場合、AI を活用することでその人のパフォーマンスがどれだけ最大化されるかを見ることに舵を切っていくと考えている。そうするために、AI を使用しなかった場合のパフォーマンスを測り、その上で AI を使用することでどこまでできたかパフォーマンスを見ることになる。ここで大切なのは、AI をどのように使用したかを説明させたり、プロンプトを提出させたりすることである。(A 社)

→ いかに AI から適切な回答をさせ、その中から正しい回答を選択する力をつけていくことが今後重要になると理解した。(インタビュアー)

→ 現状は、AI の利用を禁止していることが多いという現状であるが、それには限界がある。今後は、「禁止」より「条件付きで許容＋AI 利用の申告」が現実的と考えている。(A 社)

→ ただ、AI を利用するには課金が必要になるため、経済格差には配慮する必要があると考えている。(A 社)

② 生徒の思考力への影響

・ 生成 AI サービスの普及後、生徒の思考力への影響が確認されているか。外部公開されているレポートや貴社調査結果などで、確認されている事項がありましたらご教示いただきたい。(インタビュアー)

→ AI 固有で顕著に表れることとしては、ものを作ることが簡単になった。ボタンをクリックするだけで、クリエイティブな作業が自動化できるようになった。その結果、自分で試行錯誤した経験がある人にとっては、見極める、吟味できる力を発揮できる。ボタンをクリックする経験ばかりでは、そのような力は養われないリスクはあると考える。よって、実体験は非常に大切である。相手の状況を想定し選択して、価値判断することを伴わない思考力は、価値を持たないことが AI の登場により明確になったと考えている。（A 社）

・ 教育に生成 AI を利用することについて、ポジティブな側面とネガティブな側面の両面でどのような影響があると考えるか。（インタビュアー）
→ ポジティブな側面としては、これまで能力を発揮しにくかったマイノリティの人々が力を発揮しやすくなる。また、即時にフィードバックを得られることや、探究的な学習活動を進めやすくなることも挙げられる。ネガティブな側面は、すべてのテクノロジーで言えることだが、格差を縮小しようとするテクノロジーは、逆に格差を拡大させている。テクノロジーを使う意欲・意思があり続けられる人がどんどん先に進み、使う意欲・意思がない人が置いていかれ格差が拡大するのが大きな問題と考えている。また、AI に依存し自身で考えなくなることも考えられる。（A 社）

③ 今後の展望

・ 教育の効果を最大化するために、生成 AI をどのように活用することが有効だと考えるか。活用場面、活用方法等の観点でご教示いただきたい。（インタビュアー）
→ 根本的には、管理から自律に向かうことが大切であると考えている。国が教育委員会を、教育委員会が学校を、学校が先生を、先生が生徒を管理するために使うツールは 20 世紀の発想である。21 世紀の学びは、一人ひとりが自立してモチベーションを持ちながら学ぶことが大切である。AI をアレンジして自分に合ったようにカスタマイズすることが大事で、そのためには一人ひとりが自分を理解し、主体的に行動する必要がある。（A 社）

・ 上記のように教育の効果を最大化するために、どのような課題があり、今後どのような対策を講じるべきだと考えるか。（インタビュアー）
→ 一番は、「意欲」と考えている。学校で AI を使用したことで起こったこととして、勉強ができる学校で AI を自由に使って良いと話すと、何を作ったらよいか質問する生徒が多かった。何をすれば評価されるかを軸に行動する生徒が、真面目な生徒ほど多い。自分の中にモチベーションを持つことが大切である。（A 社）
→ これからは、意識を変えていく必要があると理解した。（インタビュアー）
→ 日本では、科学的な正しさよりもみんなが言っていることが正しい。みんなに褒めら

れたり、評価されたりすることが社会に根付いている。その中にある教育において、日本人はサイエンスの素養はあるが生かせていない。また、論理的思考力があるにもかかわらず、みんなが言っていることが正しいとすり替えてしまうところがあるため、そこは変えていく必要があると考える。みんなのことは把握しながらも、こちらのほうが正しい、良いと冷静に考えることも大切と考えている。(A 社)

→ 課題と対策として、表 6 のようなものがあげられるが、教員が見張るだけではなくメカニズムがないと広がっていかないため、思考が停止してしまい、AI の使用禁止となるのではないかと考えている。(A 社)

表 6 教育への AI 活用に関する課題と対策例

課題	対策例
評価の形骸化	プロセス重視ルーブリック（評価基準表）を作成する、回答の口頭確認を行う
情報の信頼性	出典を必須する、検証手順のテンプレートを作成
依存	カリキュラム内で AI 利用不可と可の課題の配分を明示する
教員負担	ルーブリックの共通テンプレートを用意する、少数教科から段階的に導入する
公平性	校内端末の貸出の仕組みを整備する
安全	児童生徒の個人データは持ち込まない
著作権	二次利用範囲やクレジットの方針を明確にする

B 大学（大学）

① 生成 AI の利用実態と学習方法の変化

・ 大学において、生徒による生成 AI の利用はどの程度進んでいるか。また、講師による講義や研究・事務作業等への生成 AI の利用はどの程度進んでいるか。(インタビュアー)

→ 大学としてポリシーを明確にしている。Google Gemini を法人契約し、有料版を全学生と教職員が利用できるようになった。業務個人情報や研究データを扱う場合、大学が提供している Gemini を使って業務ができるようになった。学生の学習に関しては、英語の授業で翻訳機能として使うのは語学学習として意味がないが、ルーターの Config ファイルを作る際やプログラミング学習時にエラーが出た場合に生成 AI を活

用しエラー箇所を見つけるような用途は有用だと考えており、ティーチングアシスタントのような役割で役立っている。また、一般論として卒論・レポートを書く際に AI の生成物をそのまま提出することが危惧されており、実態として私の授業でも学生が、AI が生成したと考えられる提出物を多数提出していると思われる。そうした状況から学生に AI の利用をヒアリングすると、積極的に使っていることを発表するよりは、少し後ろめたさを持って使っている印象がある。さらに、状況は日々変化し半年前と今では実態が変化していて、利用状況が異なっている。2025 年 6 月くらいに Gemini のキャンペーンがあり日本の大学生向けに 2 年間無償で契約できることになりかなり普及した気がする。そのような経緯があり、当大学では法人契約に辿り着いている。卒論・修論や学会向けの論文指導に関して、生成 AI で作成したと思われる文章が一部含まれるものが出てくることがあり、学生にヒアリングしたところ、「自分が記載した文章をブラッシュアップするように使用したがその際に意図しない記述が混入してしまった」とのことであった。(B 大学)

→ 全学生に提供しているのか。(インタビュアー)

→ あっている。(B 大学)

→ 利用に関して、ユーザーのリテラシーを高める活動はあるのか。(インタビュアー)

→ すべての学生に統一して何かを教えるということはしていない。個別の授業の先生の範囲で進めていて、学校として大きな指針はない。教授陣の間でベストプラクティスが出回っているような状況である。なお、ある学部で配布資料内に目視できないバックグラウンドと同色の嘘の情報を入れた PDF を配布し、その PDF を生成 AI に読み込ませレポートを作成した人は検知できる仕組みを用いて啓蒙活動をおこなった話は聞いたことがある。(B 大学)

・ 生成 AI は大学のどのような場面で活用されているか（自学（課題への取り組み等）への利用も含む)。もしくはどのような場面で活用することが効果的であると考えるか。(インタビュアー)

→ SNS のマーケティングに使う画像や動画の課題を出した際に、生成 AI を活用することで作品のクオリティが上がっている。3D プリンタが少し前に流行りプロトタイピングをする実習的な授業がおこなわれるようになったが、生成 AI により動画や音楽なども含めたあらゆる作品でプロトタイプを作ってコンセプトを形にすることを教育に取り入れられるようになった。また、オンラインで授業をする際、Google フォームで出席確認を行っていることがある。講演型の授業で、感想を書かせることにより出席確認を行うことがあったが、300 人中 2 人程度、実際には話していない内容がレポートに記載され生成 AI にゲスト講演者の講演のサマリーを書かせていることが推測されるものがあった。(B 大学)

・ 生成 AI の活用によって、学習の方法に変化は発生しているか。(インタビュアー)

→ ネットワーク構築の専門授業において、エラーが発生した際に、学生アシスタントが対応していた業務を生成 AI が代わりに対応している。また資格学習させる際、自分が何を理解できていないのかを見出して、提案させている。さらに、研究においては、文献調査やアイデア出しの壁打ち役として活用している。加えて、論文の読み方が変わり、例えば 10 個ぐらいの論文を AI に要約させ、要約文を確認し、適切な論文を自身で深く読むようになっている。これにより、自身の目的に即した論文を特定するスピードが上がっている。なお、10 個の論文を見つけることも AI に頼んでいることもある。(B 大学)

・ 生成 AI を使いこなす（プロンプトの工夫等）により、学生間でアウトプットの質に差が見られるといった事象は確認されているか。(インタビュアー)

→ 学生が直接プロンプトにどのようなものを入れているか把握できないが、AI の使い分け、有料サービスの利用有無で差があると思っている。また、論文の研究をする際に、AI アシスタントを使うケースがある。このアシスタントは AI にどのようなプロンプトを入力すればよいかのサポートを行ってくれるものである。また、アウトプットのクオリティを指定するなど、AI に対して明示的に指示することが重要である。(B 大学)

・ 課題等で生成 AI を利用しない前提の場面もあると考えているが、自宅等学生が見えない場面での生成 AI 利用の統制をどのように図っているか。もしくは生成 AI の利用前提で課題を考えられているか。(インタビュアー)

→ 授業内容に依存する。スキルを身に着けようとするときに、自分でやらないとスキルが付かないもので生成 AI を使われてしまうと困ってしまう。まだ、翻訳ソフトが搭載されているエディタも存在する状況で、先生が英語の記載方法を確認することに意味があるのかは疑問に思う。むしろ生成 AI の利用を前提として、文章全体を見た方がよいのではと考えている。日本人的にはスキルが身につかないとダメという話になるが、一方で生成 AI を利用する前提の考え方に移行した方が良いのではと考える。つまり道具を使いこなすことが重要であると考える。(B 大学)

・ 生成 AI の普及前後で教育の方法に変化はあったか。(インタビュアー)

→ 今までソーシャルコンテンツ制作という授業では、画像やビデオの編集など、ツールの初歩的な利用方法を教えることが多かった。それが今では生成 AI により簡単に Mock を作ることができるようになっている。生成 AI の活用を前提とすることで、教育の内容は「何を教えるか」へとシフトしてきている。ソーシャルメディア上のコンテンツで人を感動させたり、行動変容を促したりするにはどうすればよいかといった、本来注力すべき教育分野に取り組めるようになってきていると考えられる。(B 大学)

② 生徒の思考力への影響

・ 生成 AI サービスの普及後、生徒の思考力への影響が確認されているか。（インタビュアー）

→ 本質的なところの議論ができるようになった。ファクトチェックのプロセスは学習者が AI に教えることに近いことをするため、学習者にとって勉強になる側面もある。教える方が内容をより知る必要がある。AI が出力したものを何も考えず使用する人は能力が向上しないが、考える人は能力が向上していくと想定される。生成 AI に質問する際は、AI に伝わるように言語化して説明する必要がある。人に正しく聞く力が必要になってきているので、ポジティブな影響が出ている。（B 大学）

・ 教育に生成 AI を利用することについて、ネガティブな側面でどのような影響があると考えるか。また、どのような対策でケアすることが考えられるか。（インタビュアー）

→ 生成 AI の出力を自身で理解していないケースが考えられる。例えば学生が研究室の進捗報告の資料に生成 AI を利用することがある。正しく理解しているかどうかは質問を通して分かるため、質問して理解度を確認することでケアすることが考えられる。（B 大学）

・ 生成 AI の利用のネガティブな側面として、批判的思考力の低下や論理的思考力の低下といった事象（もしくは予兆）は確認されているか。（インタビュアー）

→ エビデンスのある形で提示できるものはない。ケーススタディのような個別事例だが、授業内で学生に 15 分間で回答を作成させたとき、その回答は明らかに 15 分で作成できないほど高い品質のものが提出されることがある。こういったケースはすぐ分かるが、個別事例の話となる。（B 大学）

③ 今後の展望

・ 教育の効果を最大化するために、生成 AI をどのように活用することが有効だと考えるか？活用場面、活用方法等の観点でご教示ください。（インタビュアー）

→ 現場では、ハンズオントレーニングやプログラミングの授業で間違いを探して修正するアシスタントとしては、人間よりも効率的である。抽象度の高い話だと、大学内で AI センターができ、教育でどのように AI を利用するか等を議論している。博士課程程の学生レベルの議論が現状の生成 AI とできると言及する教員もいる。また、学生が AI を用いた論文作成アシスタントを活用することで、論文の基本的な構成や誤字脱字の修正等が済んだある程度体裁の整ったドラフトを持ってきてくれるため、本質的な議論に使う時間が増える。（B 大学）

・ 教育の効果を最大化するために、どのような課題があり、今後どのような対策を講じるべきだと考えるか？（インタビュアー）

→ 半年前は有料サービスを利用できる人、できない人で差がつくということがあった。我々の大学では解決したが、有料サービスを契約している大学としていない大学で差が出てくると思う。また、初等教育や中等教育で生成 AI を触れてきた学生とそうでない学生で差がつくのではないかと思われる。さらに、日本で AI の教育現場での利用をネガティブだと決めつけるような人が出てくると、日本と海外で差が出てくる可能性がある。計算力を上げる、漢字を書けるようになるといった手段の教育に力点を置くことで、本質の検討に関する能力開発が重点的に行われなくなることを懸念している。加えて、政策という観点では AI が教育現場で使えるようにすることが大事だと思っている。教育無償化の話で PC 配布等が挙げられるが、その中に AI も含まれていくべきだと考えられる。(B 大学)

C 社（情報セキュリティベンダー）

① 生成 AI を悪用したサイバー攻撃の実態と傾向

・ 現在、生成 AI が悪用されているサイバー攻撃の数や質、および具体的な手口やトレンドは何か。(インタビュアー)

→ スミッシングにおいて文面の作成自動化で利用されていることは昨年においても確認しており、当たり前になってきているようである。ブロッキング回避が目的と思われるが効果的かどうかは疑問。ランサムウェア生成や不正ログインの悪用においては国内報道があったことを把握している。(C 社)

→ 「スミッシングにおいて文面の作成自動化で利用されていることは昨年においても確認しており」について、貴社として具体的にどのようなことが行われていることを観測しているのか補足いただきたい。(インタビュアー)

→ スミッシングに関しては、XLoader（別名 MoqHao）というアンドロイド端末用のマルウェアがある。これが不在通知を装ったスミッシングと言われており、同じ意味合いの固定文字の文面で言い回しを少し変えたものを大量に生成していることを確認している。なぜ生成 AI を用いたものかと推測できるかというと、文面の最後に生成 AI の補足説明文が加わったものがスミッシングマルウェアとしてばら撒かれていることを観測しているからである。(C 社)

→ XLoader は、SMS の内容を転送したりすることで、SMS OTP を突破することを可能にしたりするアンドロイド端末用のマルウェアと認識している。目的は、ブロッキング回避と考えて良いか。それとも、エンドユーザーを騙しやすくするためなど他にあるのか。(インタビュアー)

→ 文意は変わっておらず、人を騙すための文面の工夫はされていない。そのような文面を大量に生成しているため、ブロッキング回避が目的と思われる。どれだけ回避でき

ているかは不明である。(C 社)

→ XLoader で文意が異なるものは確認されていない認識を持ったが、その理解で良いか。(インタビュアー)

→ いくつか異なるパターンはあるが、それらが生成 AI によって生成されたものではないと考える。以前からあったものと考えている。(C 社)

→ 不在通知の言い回しは変えず、ブロッキングを回避することが主な目的であると理解した。(インタビュアー)

→ 生成 AI を使う前からある不在通知を装ったものは、アンドロイド端末でアクセスした場合はマルウェアに感染させるサイトに誘導し、iPhone 端末でアクセスする場合は Apple 社を騙るフィッシングサイトへ誘導している。
また、事業者が SMS のブロッキングを文面完全一致で行っているのか、URL 単位で行っているのかは分かりかねるが、文面完全一致で行っている場合は、多少言葉の入れ替えなどがあっても意味が通じるため、完全一致ではすり抜けやすい。(C 社)

→ APT 攻撃においては、ウクライナの CERT（CERT-UA）が Pawn Storm（APT28）に関連する可能性を指摘した事例として、LLM（大規模言語モデル）を利用するマルウェア「LAMEHUG」が報告されている。このマルウェアは、感染端末で情報やファイルを収集するためのコマンドを LLM にその場で生成させるものであった。ただし、生成されるコマンドは動的に作る必然性のない単純なものであり、検知回避が目的ならば、コマンドをハードコードして難読化する方が効率的である。このことから、このマルウェアはまだ概念実証（PoC）の段階にある可能性が高い。また、弊社で直接的な確証を得ることは困難であるが、フィッシングメールやデコイ（おとり）文書の作成において、対象国の言語や文化を自然に反映したコンテンツを生成するために AI が利用されているとみられる。さらに、サイバー攻撃とは少し異なるが、影響力工作（Information Operation）の領域では、コンテンツの大量生成のために AI が活発に利用されている。こうしたコンテンツには、生成 AI が出力する特有の冗長な表現などがそのまま含まれていることがあり、それが AI 悪用の痕跡となる場合がある。(C 社)

→ 「PromptLock」というマルウェアと「LAMEHUG」というマルウェアの違いについて、貴社としてどのように捉えているのか。(インタビュアー)

→ 双方、生成 AI の使い方の目的は、類似している。PromptLock は、スクリプトを作成するプロンプトを発行し、LAMEHUG は情報収集のコマンドを発行している。異なる点は、PromptLock はインターネットに接続することなくローカル環境で動作するが、LAMEHUG は、パブリック AI である Hugging Face を使用している点である。マルウェアの検体内部にサービスを利用するための 282 個の API キーが埋め込まれており、その中で利用できる API キーを使いパブリック AI に対してクエリーを投入するという違いがある。(C 社)

→ LAMEHUG による不正な通信を観測できるということか。(インタビュアー)

→ その通りである。パブリック AI への通信、プロンプトや応答を監視するセキュリティソリューションを導入し、Hugging Face を監視対象に含めていれば、不正な通信を防ぐことは可能である。Hugging Face をブロックすると影響が出るわけではない。生成 AI の通信を適切に監視することが大切であることが、事例から分かるところである。（C 社）

→ PromptLock はローカル環境で動作するとのことだが、動作するための条件はあるのか。（インタビュアー）

→ 正確に把握できているわけではないが、gpt-oss-20b というモデルを使用している情報は把握している。このモデルが動作することが条件と考える。（C 社）

→ この 2 つの新種のマルウェアは、スクリプトをその場で生成し必要なコマンドを投入していくもので、コマンドをローカル環境で生成するか、パブリック AI に接続し生成するかの差異があり、動作に多少の違いが出てきていると理解した。（インタビュアー）

・ 近年、攻撃者が生成 AI を利用することによって変化した「攻撃の高度化」や「自動化」の特徴は何か。（インタビュアー）

→ そこまで劇的に変わった印象はない。もちろん AI 生成による画像や文面はたとえば詐欺では当たり前のように使われているが、細かいところは人間系で補っている印象。特に SNS 型投資詐欺やロマンス詐欺では、まだ人間が受け答えを担当している部分が残っており、リアルタイムフィッシングにおいても半自動化で人間が対応していると思っている。（C 社）

→ APT 攻撃のオペレーション（マルウェアを含む）において、現状では AI が活発に利用されているという顕著な傾向は見られない。マルウェア自体も、AI の登場によって著しく高度化・多様化したという状況は確認されていない。APT 攻撃は、不特定多数ではなく、価値の高い特定のターゲットに対して秘匿性を保ちつつ実行される。そのため、攻撃の実行段階で AI を利用する必要性は限定的であると推測される。ただし、言語や文化の壁を越えるハードルは着実に低下している。従来見られたような、機械翻訳に起因する「明らかに不自然な」日本語の標的型攻撃メールは減少傾向にある。（C 社）

→ APT 攻撃は従来から個別にカスタマイズしたマルウェアが多く、AI は量産性、効率化を重視しており、生成 AI を使用するのは、あくまで攻撃の準備段階で使用され、限定的な使用と理解している。（インタビュアー）

→ どちらもある。マルウェアを作成する段階で AI を使っているかもしれないが、その点については把握していない。実際のオペレーションでラテラルムーブメントする際に AI を使って自動化することも可能であると思われるが、APT 攻撃は AI を使い幅広く攻撃するものではないため、攻撃対象がミスするリスクをとるより、確実に組織の持

っているプロシージャに沿ってオペレーションする方を選択すると個人的には考えている。(C 社)

→ まだまだ人が仲介して、AI 任せになっていないと理解した。(インタビュアー)

→ これは、企業のシステムに不正アクセスし横展開する際の Post-Exploitation(システム侵入後の活動)の話となる。人が仲介しているソーシャルエンジニアリングにおいては、日本語への翻訳、相手の文化に合わせた文面を作成するところで AI を使用しているかもしれないが、Influence Operation のように大量にコンテンツを生成し配信する使い方はおそらくしていないため、生成 AI で作られた余計な文章は入っていない。ソーシャルエンジニアリングでも同様に使用していると思われる。(C 社)

→ 一方、裏側では文面生成で AI を使っている可能性はある。サイバー犯罪だと生成 AI っぽさがやりとりで残ったりしている。(C 社)

→ 不用意な痕跡の残さないように裏側で使用していると理解した。(インタビュアー)

→ AI の挙動が検知される可能性や AI が生成する結果が誤っている可能性もあるため、そこまで使用されてはいないと考える。(C 社)

→ 「PromptLock」や「LAMEHUG」による感染後のラテラルムーブメントでは、人が介在してくるという認識を持たれている理解で良いか。(インタビュアー)

→ ラテラルムーブメントについては、従来の自動化ツールは使われていると思われるが、それ以外で AI を特に使うということはない。(C 社)

・ 攻撃対象として特にリスクが高まっている業種や規模、地域的傾向はあるか。(インタビュアー)

→ 情報を持ち合わせていない。推測としては言語の壁は確実に下がっているという印象。(C 社)

→ 生成 AI の活用により、従来は言語や文化の壁が高いとされてきた日本のような地域への攻撃の障壁は、以前よりも低下しているとみられる。また、マルウェア自体が動的に LLM を利用するような攻撃においては、自社のネットワークから外部の生成 AI サービスへの通信を適切に管理・制限できていない組織や個人が、より標的となりやすくなる可能性がある。(C 社)

・ 日本のように言語による攻撃の障壁が高い国があれば教えていただきたい。(インタビュアー)

→ 他の国については把握していない。(C 社)

→ 日本は海外に比べると言語の壁が高いと考えられるのか。(インタビュアー)

→ 言語の壁のほか、ビジネスメールの文頭の挨拶文などは従来の機械翻訳では難しい部分はあったが、このような慣習の部分は AI を使うことで突破することができると考える。(C 社)

・　生成 AI を悪用したサイバー攻撃であると推定するにはどのような方法があるか。（インタビュアー）

→ 生成 AI ではテキスト、音声、動画、イメージ、コード生成がある。それぞれ検知できるサービスが異なっており、その精度も異なる。特化したサービスが複数登場してきている。サイバー攻撃の種類によって推定度合いが異なる。たとえば詐欺ならプロフィール画像が AI 生成というのは見かけるしそれ自体は検知サービスを使えば判定できるとは思う。ただし生成 AI を使用することは一般的にも当たり前になってきているので、それだけで詐欺や悪性との断定が難しい。また、国内逮捕された事例で正規サービス利用というよりジェイルブレイクしたサービスを利用しているものもある。加えて、複数サービスのコードを組み合わせているという事例も把握しており、これにより検知が難しくなるのではないかと考えている。（C 社）

→ 国内の逮捕事例に関して、マルウェアが AI で生成されたことがなぜ分かったのか。（インタビュアー）

→ マルウェアの動作に LLM が利用されているかを特定する方法として、リバースエンジニアリングによる解析が挙げられる。デコイの生成に AI が用いられたかを推定する方法には、以下の手法がある。（C 社）

- LLM の応答に含まれがちな定型句や冗長な表現が混入していないかを確認する。
- 「Gen AI Detector」のような検知ツールを利用し、テキストの複雑さ、文章の均一性、語彙の多様性、特定表現の反復などを分析する。

・　攻撃者側が利用する犯罪用生成 AI ツール（WormGPT など）の開発・流通状況はどのようなものか。（インタビュアー）

→ WormGPT がサービスとして稼働していることはウェブサイトやテレグラム等をみれば把握できる。また他の類似サービスが乱立していることも把握している。ただし、実際に利用したわけではないので、どこまでの精度なのか実際に利用した場合がどうか把握していない。このようなサービスでは ChatGPT などの正規サービスをラップしただけのものもあると聞いている。（C 社）

→ 留意すべきは、攻撃者は必ずしも犯罪専用のツールのみを利用するわけではなく、ChatGPT や Gemini といった正規の汎用サービスも悪用する点である。前述の「LAMEHUG」の事例では、正規のプラットフォームである Hugging Face 上でホストされ、無料で利用可能な生成 AI の API が悪用されていた。（C 社）

② 生成 AI を悪用したサイバー攻撃に対する対策

・　生成 AI による脅威に対応するため、従来のセキュリティ対策と異なる点はどこにあるか。（インタビュアー）

→ 生成 AI による脅威への対応には、従来の手法に加え、以下の点が重要となる。（C 社）
●ディープフェイクやパーソナライズされたフィッシングなど、巧妙化する攻撃への対策
人間には見抜くことが困難な偽の映像や音声を検知するため、映像の微細な色の変化、不自然な瞬き、音声に含まれる合成ノイズなどを AI で分析・検知する技術の導入が不可欠である。
●攻撃の「量」と「スピード」の増大への対策
既知の脅威情報を基にしたシグネチャベースの防御だけでは不十分である。AI を活用し、ネットワークや端末における「平時と異なる振る舞い」をリアルタイムに検知し、未知のマルウェアであっても自動で隔離・遮断する仕組み（EDR/XDR など）の導入がより一層重要となる。

③ 生成 AI を悪用したサイバー攻撃の対策における社会的ガバナンスと連携・啓発

・ 生成 AI のサイバー攻撃に対して、貴社のようなサイバーセキュリティ企業のみでなく、ユーザー企業（一般利用者も含む）に求められる注意点やリテラシー向上策はどのようなものがあるか。従来の AI が普及する前とどのような差分があるか。（インタビュアー）

→ 一般利用者においては従来からの姿勢と変わらない。例えばフィッシング詐欺においては既に見分けるのは不可能であり、提示された URL からアクセスするのではなく、どんな時も正規アプリやブックマークからアクセスしてログインすることを習慣づけることが重要である。警察官を騙る電話詐欺でも、自ら調べた警察署の電話番号にかけなおすことを習慣づけることが適切。見破れない前提で行動することが、ますます重要になる。質というより生成 AI により仕掛ける側が効率化し量が増加することが予想される。また、詐欺は見破ろうとしないことが大切である。相手から提示されたものが正規かどうかに関わらず自ら確認する習慣が一般利用者の視点からは、一番の防御となる。（C 社）

→ ユーザー企業においては、従業員による不適切な情報漏洩を防止するだけでなく、LLM を利用するマルウェアからの防御という観点からも、生成 AI サービスへの通信を適切に制限・監視することが有効なリスク低減策となり得る。ただし、個々の利用者レベルでの対策は極めて困難である。生成 AI によって攻撃が巧妙化していく現代においては、個人のリテラシー向上のみで脅威を防ぐことには限界があると認識すべきである。（C 社）

・ 生成 AI を悪用したサイバー攻撃に対して、産学官や外部機関等との情報共有・連携の中で、生成 AI に特化した連携で特に重要だと感じているテーマや課題があれば教えていただきたい。（インタビュアー）

→ 脅威の種類に応じて、その利用形態や精度が異なるので、現実的な脅威分野に分けて、国内とグローバルで分析していくことは変わらないと考えている。従来の分野に生成 AI の要素が加わる。一方、生成 AI サービス・それを活用したシステム自体を攻撃対象にした場合に攻撃手法や攻撃領域が従来のものと異なるので、ペンテスト分野においては特化したものとなりうるし、起こりうる影響を考えるうえでも専門的知見が必要になると思っている（C 社）

→ また、以下について、特に重要と考えている。（C 社）

●脅威情報の共有高度化

マルウェアの動作がより動的になるため、IP アドレスやハッシュ値といった従来の IoC（Indicator of Compromise）の共有だけでは不十分である。悪用されたプロンプトの内容や生成 AI モデルの種類といった、新たな脅威インテリジェンスを迅速に共有する枠組みが必要である。

●コンテンツの信頼性確保（デジタル来歴）

画像や動画が「いつ、誰が、どのように作成・編集したか」という来歴情報を記録・検証できる技術（例：C2PA）を社会インフラとして普及・定着させることが重要である。これには、報道機関、プラットフォーマー、デバイスメーカー、政府機関などの広範な連携が求められる。

●真正性データベースの構築

政府や企業の公式サイト、公式発表の動画・音声などを「真正な情報」として登録するデータベースを構築し、市民がファクトチェックのために参照できる仕組み作りが必要である。

●法整備の検討

ディープフェイクによる詐欺や名誉毀損といった新たな脅威に対し、現行法でどこまで対応可能か、どのような法改正や新たな立法が必要かを、法律家、技術者、政策担当者が連携して速やかに検討する必要がある。

D 社（情報セキュリティベンダー）

① 生成 AI を悪用したサイバー攻撃の実態と傾向

・ 現在、生成 AI が悪用されているサイバー攻撃の数や質、および具体的な手口やトレンドは何か。（インタビュアー）

→ サイバー攻撃の数や質は、劇的に向上している。メールセキュリティの分野から見た場合、フィッシングメールの数が増加している。メール添付のファイルでマルウェア感染させるものではなく、相手をソーシャルエンジニアリングで騙して正規のサイトを騙ったサイトに誘導し、ID とパスワードを入力させてログインさせる。バックグラウンドでセッションキーを窃取し、多要素認証を実施していても、それを迂回して攻

撃対象の組織に不正アクセスする手法の攻撃がかなり増加している。現在の量は、2023 年 1 年間の平均値の 7〜8 倍である。ほぼ日本をターゲットとしているところが特徴である。先月 8 月の統計で新種のメール脅威のうち、79.6%が日本をターゲットとしている。今までは言語の壁があり、日本語の言い回しなどからフィッシングメールだとすぐに見分けることができたが、攻撃者が AI を使用するようになったことで自然な日本語の文面を大量に様々なパターンを生成することができるようになった。これが攻撃の増加につながっていると考えている。その他 AI が関与しているものとしては、ボイスフィッシング（電話による詐欺）がある。当社の顧客から確認したのだが、社長の声を AI に学習させ、社長の声をまねした声でしかも社長の電話番号で電話がかかってくるのが特徴である。声をまねさせる、学習した声で自由に発言させるほかに、AI ツールの中にはソーシャルエンジニアリングを自動で行うものもある。今までは攻撃者が人件費や時間をかけてやっていたことが、AI により自動で攻撃できてしまうところが影響していると思われる。（D 社）

→ メールフィッシングでは、企業の従業員を狙うもの、決済サービスのユーザーを狙うもの、どちらが増加しているのか。それとも両方増加しているのか。（インタビュアー）

→ 両方増加している。（D 社）

→ クレデンシャルフィッシング（認証情報窃取）には 3 つの攻撃パターンがある。企業向けでは Office 365、Google Suite、DocuSign などの認証情報を窃取するもの、個人向けではクレジットカード、PayPay、Amazon、証券口座などの認証情報を窃取するもの、どちらとも区別がつかないものに分けられるが、いずれも増加している。90%以上が認証情報窃取となる。2025 年の 1〜8 月統計では、全体の新種のメール脅威が約 52 億 7 千 6 百万通程度である。87%フィッシングメールで内訳としては 31%企業向け、31%が個人向け、25%が企業向けか個人向けか判断できないものとなる。（D 社）

→ 対策としては、文面が多様になっているため、文面マッチングでのブロッキングは難しいと認識しているが、いかがか。（インタビュアー）

→ 攻撃キャンペーンごとに文面が変わり、キャンペーンの量も多くなっている。（D 社）

・ 貴社のブログ記事を拝見し、攻撃対象として特にリスクが高まっている地域として、日本が挙げられている。生成 AI により流暢な日本語が作成可能になったことが一因とされているが、他に要因はあるか。（インタビュアー）

→ 日本は、言語の壁がありセキュリティ対策が進んでいなかった。言語の壁を越えた今では非常に攻撃しやすい状況にある。（D 社）

→ 日本の企業の対策がグローバル企業よりも遅れているとのことだが、具体的にどのような技術の導入が遅れているのかわかればお伺いしたい。（インタビュアー）

→ 詐欺メール対策であれば、なりすましメールをブロックするための DMARC、送信元

のブランドロゴを表示する BIMI、電話番号の詐称対策の STIR/SHAKEN 等が挙げられる。(D 社)

・ 生成 AI を悪用したボイスフィッシングについて、他に事例があれば、詳細をご教示いただきたい。(インタビュアー)
→ 他には 5 社への攻撃があり、業種としては製造業が多かった。(D 社)
→ いずれも社長、役員を名乗ったものか。(インタビュアー)
→ ほぼ同じやり方で狙っている。(D 社)
→ 銀行からの振り込みがうまくいかなかったため、至急振り込み処理してもらいたいというようなフィッシングは観測されているか。(インタビュアー)
→ ボイスフィッシングの電話があった先について、どういった結果に終わったのかお伺いしたい。(インタビュアー)
→ 実際に送金した会社が東南アジアで 1 社あった。日本企業では社長の一言で大きなお金は動かせないので、確認する中で詐欺であることが判明したケースが多い。海外の小さな法人が危ないとのこと。(D 社)

・ 生成 AI を悪用したサイバー攻撃であると推定するにはどのような方法があるか。(インタビュアー)
→ マルウェアのソースコードであれば、AI で生成したことが分かるコメントがあったりすることはある。AI によりサイバー攻撃が高度化して量が多くなっているが、守り方は同じである。守りにも AI を使用しなければ対応に手が回らない状況である。(D 社)

② 生成 AI を悪用したサイバー攻撃に対する対策
・ 生成 AI による脅威に対応するため、従来のセキュリティ対策と異なる点はどこにあるか。また、技術的な対策と人的・組織的対策にはどのようなものが挙げられるか。(インタビュアー)
→ ボリュームとスピードについていける対策となっているかである。AI による検知など、防御でも AI を活用することが重要である。(D 社)
→ 技術的、人的対策で何か言えることはあるか。日本は他国と同一レベルの企業と比較して対策は遅れている点があるとのことで、技術的な対策としてはグローバルスタンダードレベルとなっていないと認識しているが、その認識で合っているか。(インタビュアー)
→ 合っている。(D 社)
→ 組織的なところで、気を付けなければいけない点があれば伺いたい。(インタビュアー)
→ AI には、AI for Security、Security for AI の両方がある。AI エージェントは人に与えて

いる権限と同じように、それがどれだけデータに対するアクセス権を持つのかが重要となる。データへの認可は、中央集権的に管理することはできない。各業務分野にアクセス付与の認可を任せるようになると考えている。認可する担当者にサイバーセキュリティのナレッジ、経験がない場合は、すべて認可するような対応となる懸念がある。AI を使いこなす人がセキュリティを意識したデータへの認可権限を与えなければいけない。そこをいかにトレーニングするかが課題である。(D 社)

③ 生成 AI を悪用したサイバー攻撃の対策における社会的ガバナンスと連携・啓発

・ 今の状況を踏まえ、作ったら良い AI についてのガイドラインなどがあればコメント頂きたい。(インタビュアー)

→ AI の利活用を促進するようなものであることが望ましいと考える。(D 社)

・ 生成 AI を悪用したサイバー攻撃に対して、産学官や外部機関等との情報共有・連携の中で、生成 AI に特化した連携で特に重要だと感じているテーマや課題があればご教示いただきたい。(インタビュアー)

→ 攻撃の情報共有が防御を高めることができるので、攻撃の情報共有が重要であると考える。システム連携で情報共有できるようにするとよりよいと考える。(D 社)

・ 上記でご回答いただいたもの以外に、生成 AI を悪用したサイバー攻撃に関して重要な論点などございましたらご教示いただきたい。(インタビュアー)

→ それ以外では本人確認（eKYC 等）が重要と考えている。例えば電話番号が正しくても本当に本人から電話されているかの確認は容易ではなく、担保していくことが重要であると考える。(D 社)

<u>E 社（児童保護団体）</u>

① 生成 AI による児童性的被害の現状把握

・ AI 由来の児童に関するわいせつ物の事例数や国内相談件数はどのようなものか。(インタビュアー)

→ 件数は、相談窓口を設けている団体やネットパトロールを行っている団体などから聞くのがよい。海外では、米国の NCMEC や英国の IWF のレポートから入手可能。(E 社)

→ 日本国内の事例数や相談件数について公開しているところはあるか。(インタビュアー)

→ AI についての事例数や相談件数を出しているところはない。AI 由来かどうかの判断が困難なため、統計的な情報を取ることは難しいと考える。(E 社)

・ 貴組織では、実在する児童、実在しない児童どちらのわいせつ物についても調査の対象とし、対策を考えているか。（インタビュアー）

→ 最終的には、AI による CSAM であろうとなかろうと、いかなる形態の CSAM もなくすべき立場にある。一気に対策できる状況ではないので、段階的に議論し進めている。（E 社）

→ どのような段階を踏むことが考えられるか、詳細をお聞きしたい。（インタビュアー）

→ 議論の最中ではあるが、CSAM の種類で考えられるものとしてⅰ）完全実写、ⅱ）部分実写、ⅲ）テキスト・音声、ⅳ）アニメ、ⅴ）すべての CSAM があり、番号順に対策優先度が高いと考えている。ⅱ）の部分実写が AI による CSAM と考えている。AI によって作られたコンテンツには、実在する児童を加工して作成されたものと実在しないものがある。そこをどう捉えるかの継続議論が必要と考えている。（E 社）

→ 完全実写から対策していくと理解した。実在する児童、実在しない児童については、議論のポイントと考える。まずは実在する児童に対する対策、その後、部分実写への対策が重要という理解で良いか。（インタビュアー）

→ その通りである。「一部実在児童」のディープフェイク（顔だけ実在児童）、ディープフェイク（体だけ実在児童）、「非実在児童」の⑤ディープフェイク（非実在児童）が AI の関わるところとして問題となっている。「一部実在児童」の場合は、優先して対応していく必要がある。AI は非実在児童を作成することができるため、現状手がつけられない。「一部実在児童」のディープフェイク（顔だけ実在児童）、ディープフェイク（体だけ実在児童）について、法改正などで対応していく必要があると考えている。また、課題として、実在性の解釈が多様である、実在性の判断が困難であるなどの点が挙げられる。（E 社）

→ 大変参考になる見解である。挙げていただいた課題があるために、法令（児童ポルノ禁止法、刑法など）が適用できない認識で良いか。（インタビュアー）

→ その通り。法律はあるが、実際に適用し運用できるかというとできない、もしくは適用が確実ではないと考えられる。相談者が法執行機関に相談に行っても難しいと告げられることもある。法律で CSAM が明確に定めていないがゆえにそうなってしまうという私見である。（E 社）

→ 「実在性の解釈が多様である」について、児童ポルノ禁止法の定義では、児童が実在していることが前提にあるという理解で良いか。（インタビュアー）

→ 日本の児童ポルノ禁止法は実在していることが大前提となる。そこが多くの国と異なるところである。（E 社）

→ 実在しない場合は、今の法律では禁止とは言えないという認識であっているか。（インタビュアー）

→ 非実在児童は完全に該当しない。曖昧となっているのが一部実在児童となる。海外に

おいては、違法となるが、日本では違法かどうかを相談窓口で判断することができないことが問題と考える。(E 社)

→ 例えば、卒業アルバムを悪用された場合に相談窓口に持って行った場合、本人と認めるかどうかは、判断する人の解釈次第となるのか。(インタビュアー)

→ 顔だけで児童ポルノ禁止法の定義に合うのか。実在する主体が顔だけでも言えるのかが、判例上はっきりしない。児童をリアル CG 等で表現した場合、実在する児童をコンピュータグラフィックによりリアルな児童を表現し直しているため、違法となると考えられる。例えば顔以外が実在する児童で顔だけ AI で作成した場合、実在するかどうか判断できない。法執行機関は、過去の判例に基づき判断するため、裁判に持っていくことを躊躇しがちとなるのが実情のようである。(E 社)

・ 刑法 175 条は、流通させることが法令違反に該当する理解で良いか。(インタビュアー)

→ わいせつ性のあるものについて、それを流通させていることがわいせつ物の頒布に該当するため違法となると考えられる。海外では CSAM を違法とする法律とわいせつ物の禁止の 2 本立てで禁止している。国によって刑罰の度合いが異なるが、わいせつ物の場合は比較的軽い刑となるため、それに抑止効果があるのかが問題と考えている。児童ポルノ禁止法のようにより重い刑で罰するのが重要と多くの国で考えていると思われる。(E 社)

・ わいせつ物の事例数、国内相談件数について増えてきているか、個人的な温度感をお聞きしたい。(インタビュアー)

→ 相談窓口をおこなっている団体の話を聞いても非常に増えてきていると聞いている。アメリカ NCMEC の最近のレポートでは、一昨年と比べ 2024 年が 100%を超える勢いで AI による CSAM が増えている。同様に日本も増えているだろうと考える。(E 社)

→ 被害児童の特性や被害件数の日本特有の特徴はあるか。(インタビュアー)

→ 各国の状況や団体の話を聞いてもそれほど違いはない。但し、日本は相談窓口が少なく、分かりにくいように思う。肌感覚となるが、日本の子供たちは誰かに相談しにくいため、統計が取りにくいのではないか。(E 社)

・ AI を悪用した児童ポルノの生成や拡散の手法にはどのような傾向があるか。(流通経路等)(質問表)

→ 国内だと SNS。海外ではダークウェブもある。詳細はネットパトロールを実施している団体が詳しいと思われる。(E 社)

② 児童保護に向けた対策と課題

・ AI による児童性的虐待から児童を守る策を検討しているか。もしくはどのような対策が有効だと考えているか。（インタビュアー）

→ 今年の 3 月に超党派ママパパ議員連盟や関係省庁と話をしており、法的な対応として、児童ポルノ禁止法の法改正（AI 由来のわいせつ物は現状の児童ポルノ禁止法では対応できないため）、もしくは新法の制定の提言している。鳥取県の条例や都道府県の青少年の育成条例の動向を見ていても、法改正をする余地があると考えている。ただ中長期的な改善方法となるため、短期的に改善できる方法で対応する必要がある。制度的な対応のほか、技術的な対応についても提言を行っている。制度的なものとしては、AI による CSAM についての専門知識が必要となる。インターポール（国際刑事警察機構）は、各加盟国の警察の中に専門機関を持つべきとの見解を出している。我々も日本がそれに足並みを揃えてもらいたいと考えている。そして、相談窓口の敷居を下げる必要がある。さらに、救済措置も十分できていない。ディープフェイクの加害者が子供の場合もあるため、今の法律で対応できるものではない。その他、子供への AI に対するリテラシーを高めることが大切で、海外において非常に力を入れている。子供たちが自分たち自身を守るために、CSAM に対する判断能力を養うことは非常に大切である。技術的な対応については、AI が CSAM を生成できないように AI に CSAM を学ばせて制限することができれば、かなり改善できるのではと考えている。（E 社）

→ 改正の提言はどのような内容か。（インタビュアー）

→ 実在性から実在みたいなもの（海外では「擬似」と呼んでいる）まで含めるべきではないかという改正の提言となる。児童ポルノは、実在性、擬似、仮想の 3 つに区分される。日本は実在性だけにとどまっているが、他の国は仮想まで含め禁止している。（E 社）

・ わいせつ物の流通の実態を防ぐための課題について、法律面以外で課題があればお聞きしたい。（インタビュアー）

→ 日本は専用 Web サイト等でわいせつ画像を作成することは違法ではないと考えられる。海外ではこの日本の現状が非常に問題となっている。AI がデバイスに入っているため簡単に作れてしまう現状がある。それが簡単に海外のマーケットにまで流通してしまうところが問題である。（E 社）

③ 対策に関する他機関との連携における課題

・ 国内外の NGO・政府機関との協力体制はどのようなものであり、また課題はどのようなものか。（インタビュアー）

→ 国家機関に働きかけをしている。また、海外の児童保護団体とも連携している。課題は、上述の通り児童ポルノ禁止法の定義である。（E 社）

・　法規制に関する提言について、国内外の先進的な対応事例を参考にしているか。参考時にはどのような観点で参考にしているか。（インタビュアー）

→　常に参考にしている。法体系やカルチャーのみならず、特に運用面を重視している。（E 社）

→　特に運用面を重視しているとのことだが、重視すべき論点があればご教示いただきたい。（インタビュアー）

→　法執行機関に訴えやすくする。また、裁判官が判断し易くすることと考える。議員立法ではない形で法律を作る場合、有識者と十分議論しながら法律にし、どのように運用するかに時間をかけているが、議員立法では、それが十分に行われないまま成立してしまうケースがある。そのため、行政官の方々は運用面で負担が高いと考える。それが顕著なのが児童ポルノ禁止法である。実務を遂行していく際にどのような問題があるか、そこをクリアするために、運用面を十分に検討しなければ、運用で一番困るのは被害児童やその保護者であるため、我々はそこを強調している。（E 社）

・　海外の運用で参考になるものがあればお伺いしたい。（インタビュアー）

→　各国では AI による CSAM は違法であると言っている。更に子供たちが被害に遭わないようにするために、様々な対応を行っている。イギリスでは、AI の開発者からその他の事業者までといった幅広いステークホルダーの規制を考えている。（E 社）

・　SNS プラットフォーム事業者などへのデータ削除要請の実行性はどの程度確保できているか。（インタビュアー）

→　TakeItDown は有効なものと見ている。TakeItDown は、本人の申告に基づきデータを削除するプラットフォームである。日本も適用しているが、画像や動画の投稿が可能なオンラインプラットフォームの事業主しか参加することはできない。参加する事業主の数がより増えていくことが望ましいと考える。（E 社）

<u>F 社（ネットパトロール団体）</u>

①　生成 AI による児童性的被害の現状把握

・　生成 AI 由来のわいせつ物に関する事例数や国内相談件数はどのようなものか。また、実在する人物に関するもの、実在しない人物に関するものの割合はどのようなものか。（インタビュアー）

→　2025 年 3 月～6 月の間に当組織が行った調査では、実在する未成年画像を加工した性的事例だけで 250 件以上のフェイクポルノを発見した。そのうち 20 件は小学生が被害者である。本調査はクローズドなコミュニティサイトや匿名掲示板を対象に行った。これは、我々の観測範囲となるが、近年フェイクポルノの数が多いと思われるプラットフォームであるためである。私たちは実在の被害者がいる事案についての通報活動

を行っているため、非実在の AI CSAM について数的な把握はしていないが、パトロールしている中で増加傾向にある。(F 社)

→ 実在する児童の画像を加工したかどうかは、実在する児童の加工していない画像と加工した画像を見比べなければ難しいように思われるが、どういったところから実在する児童の画像を加工したと判断されたのか。(インタビュアー)

→ かなり長くこのようなことに関わっているため、現段階では画像を見て、実物か実物でないかをかなりの精度で判断することができる。今回の調査は、報道機関と合同で行ったものとなる。サンプルとして 10 件抽出し、報道機関から研究機関に委託し、加工したものか、1 から生成したものかどうかを判定してもらった。その結果、判定した画像は、95%から 99%の確率で実在する人物であるという結果を得ることができた。(F 社)

・ 上記のうち、生成 AI 由来のわいせつ物の対象(児童、成人等)はどのようなものが多く、またデータの形式(画像、動画など)の割合はどのようなものか。(インタビュアー)

→ 一番多いのは芸能人、アナウンサーなど著名人の被害となる。2023 年ほどから一般の中高生の被害が増加してきている。2024 年の 9 月頃は 1 枚の着衣画像から 1 枚のわいせつ画像が生成されるパターンが多かったが、昨年の年末頃から 1 枚の着衣画像からわいせつ動画や性的な行為をする動画が作られる被害の割合が増加している。(F 社)

・ ネットパトロールを行っている中で、生成 AI が普及する前(例えば 2022 年以前)と普及した後で、ネットに流通するわいせつ物の量や質に変化があるか。(インタビュアー)

→ 性的な画像・動画を販売するプラットフォームにおいて、成人・未成年問わず AI で作成したポルノが販売されるようになっている。総合的な量に関しては、自分たちがパトロールできる量に限界がありなんとも言えない。(F 社)

→ 性的な画像・動画は、取り締まりが難しい販売サイトで販売されている理解で良いか。(インタビュアー)

→ 性的な画像・動画を販売するプラットフォームにおいて、AI で生成したポルノ画像が販売されるようになっているが、非実在のものとなる。これとは別に、依頼があった場合に実在する児童の画像等の加工を請け負う者もいる。また、メンバーシップ型のビジネスを行う者もいる。(F 社)

・ パトロールを行う中で、通報する場合はどのような手段があるか。(インタビュアー)

→ 通報する手段は、いくつかある。プラットフォームへ通報すること。もう一つは、被害者が明らかな場合は学校に連絡し、学校から被害者に連絡、被害者が警察に相談す

ること。プラットフォームへの通報について、我々は第三者となるため、通報が通る場合と通らない場合がある。また、プラットフォームがどのような基準で対処しているか見えづらい。プラットフォームへの通報で削除される場合もあるが、SNS は直ぐに新しいアカウントを作成できるため新しいアカウントで同じような投稿がなされたり、別のサーバーに投稿されたりしてしまうのでその場しのぎの対応となり、無力感を感じている。こうした事例の中には、被害画像の加工後の画像と加工前の画像の両方が投稿されている事例がある。加工後は児童の性的搾取にあたるが、着衣画像は普通の画像であるため、そのまま残ってしまう。その画像が素材として残っていれば、また悪用する人が出てくる可能性がある。そのため、削除対応について包括的な対応を行ってもらいたいと考えている。後者の学校側への通報については、全体の中で対処に至るのは１％程度となる。（F 社）

・ 2025 年の春先から SNS プラットフォームではセンシティブな投稿に対して規制し始めていると理解している。先ほどのお話からプラットフォーマー側での規制には限界があると認識したが、その認識で合っているか。（インタビュアー）

→ 体感ベースではあるが、最近ではディープフェイクに関連する用語を SNS で検索すると投稿はあるが閲覧できなくなっている実感がある。ただそのような投稿をプラットフォームが自主的に削除しているのか、アカウントが凍結されているのかはブラックボックスである。非実在の CSAM は（性器が鮮明に描写されている場合はわいせつ物に該当する可能性もあるが）日本では合法という扱いなので、判断が難しいと思われる。（F 社）

・ 生成 AI を悪用したわいせつ物の生成や拡散の手法（流通経路等）にはどのような傾向があるか。（インタビュアー）

→ 性的な画像に加工するサービスが非常に多く存在している。これらのサービスは、宣伝に協力した利用者に無料加工のポイントを付与するなどインセンティブを設けていることが多く、利用者が自発的にサービスを拡散する仕組みができており、SNS やネット掲示板にサービスへのリンクが大量に投稿されている状況が続いている。また、児童には限定されないが、今までの話は素材となる画像が 1 枚あり、それを加工するものであった。しかし、それとは別に特定人物の複数枚の画像を AI に学習させることで嗜好性を持たせそっくりな人物の生成させる被害も起きている。この場合、具体的にこの画像をもとに加工されたと言うことが言いづらく被害を立証することが難しいと思われる。（F 社）

・ 実在する人物に酷似した画像も生成されている場合、被害に遭った方の心理的・社会的影響について把握されているか。（インタビュアー）

→ データを持っている近しい人物が作成、もしくはデータの提供をしていることが推測されるため「いったい誰がこんなことを」と疑心暗鬼になる、不安感を感じている被害者が多い。氏名などの個人情報もあわせて中学生のディープフェイクポルノが投稿された事例では、投稿に性加害を実行するコメントがつき、保護者が通学の送迎を行うなどご家族にも大きな負担があった。(F 社)

→ 犯人が分かったとき近しい人物が行っていた事例があったかお聞きしたい。(インタビュアー)

→ 私が通報して加害者が分かった事例では、犯人全員が同級生であった。また、被害者が警察へ相談に行くこともあるが、刑事処罰の法律の壁があり警察で取り締まりできない場合もある。また、名誉毀損に該当し得る画像と判断された場合であっても、投稿された場所が国外の掲示板、SNS などの場合、様々な条例によって開示請求ができない場合がある。被害救済するためには自分で民事訴訟を起こし、開示請求するしかない状況である。その費用負担についても被害者側にあり、経済的負担が大きい。(F 社)

② 貴組織における対策と課題

・ 生成 AI 由来のわいせつ物の生成や流通について、どのような対策が有効だと考えているか。実在する児童、実在しない児童の間で対策が異なる場合はそれぞれご教示いただきたい。(インタビュアー)

→ 実在する児童については、被害をとりこぼさないような、実効性のある法整備と諸外国との捜査協力体制の強化が必要だと考えている。非実在児童については、非常に難しい問題だが、AI 技術の発展に伴いそもそも実在児童なのか非実在児童なのかの判断が困難になりつつある。AI CSAM の氾濫は捜査リソースを奪うもので、社会的法益の観点からディープフェイクポルノを生成するサービスの中でも CSAM が生成可能なサービスについてはアクセス制限を行う等「作らせない」ことが重要ではないか。また、CSAM が生成されてしまうサービスは、移り変わりのスピードが速い。2024 年 11 月に報道機関の取材を受けた際、その時点でどんな AI サービスがあるかの調査が行われ、50 程のサービスをリスト化したが、現在ほとんどのサービスがアクセスできない状況となっている。それは、サービスが減ったかと言うことではなく、外側を変えて新しいサービスが出てきている状況である。(F 社)

・ CSAM に限定していないが、一からプロンプトにより生成させるもので、意図せず特定の人物に近い画像が出力されてしまう事例が日本のほか他の国でも起きている。このような事例にどのように対応していけば良いのか法的な判断が難しい。加害者側がそれを意図して生成したことを証明しないと加害行為が認められない。(F 社)

・ 生成 AI 由来と思われるわいせつ物を発見した際、具体的にどのような対応フローを取っているか。（プラットフォームへの削除要請、警察への通報、被害者支援など）（インタビュアー）

→ 通報の観点では、被害者の所属がわかるものについては、学校・教育委員会・管轄の警察へ通報する。被害者の所属がわからないものは、（SNS の場合）プラットフォーム、インターネットホットライン、NCMEC への通報を行う。被害者支援の観点では、画像削除関連のサービスについての情報提供や申請の支援を行っている。（F 社）

→ 同級生による加害があると、学校教育において AI の取扱いについて授業に取り込まれていることはあるのか。（インタビュアー）

→ 私立学校では、そのようなことを実施している報道は見たことはあるが、公立校を含めそのような授業を行っているかは分かりかねる。（F 社）

・ 従来のわいせつ物対策と比べて、生成 AI 由来のコンテンツ対策で特別に工夫している点や、困難な点はあるか。（インタビュアー）

→ プラットフォームに通報・削除申請をし、生成されたヌード画像は消えたが素材として提供された元画像が消えないということがある。また、生成されたわいせつ画像の性的な部位をモザイクやスタンプでマスキングしたものを投稿、「モザイクなしを見たい人は DM」までと誘導するパターンなど公然性について判断が難しい事例があると感じている。（F 社）

③ 児童保護に向けた対策と課題

・ 法規制（規制による対策）、プラットフォーム事業者（流通対策）、AI 開発者（生成対策）などに対して、生成 AI によるわいせつ物に対する対策で期待することや課題に感じていることはあるか。（インタビュアー）

→ 法規制では、名誉棄損など MLAT（刑事共助条約）の枠組み外の罪状の場合、米国にあるプラットフォーム企業に対して警察が開示請求することが難しく、現実的には被害者は民事で解決するしかない状況で被害者の負担が大きいと感じている。諸外国との捜査協力を前提とした実効性のある法整備や条約の締結が必要と考えている。AI 開発者は、CSAM を含まないクリーンなデータセットでの学習や CSAM 生成の制限を行うことが考えられる。（F 社）

・ SNS プラットフォーム事業者など（海外事業者含む）へのデータ削除要請の実行性はどの程度確保できているか。（インタビュアー）

→ 実在児童の事例の場合、米国企業のプラットフォームであれば、著作権者が DMCA（デジタルミレニアム著作権法）で通報すれば加工の内容に関わらずかなりの精度で削除が実行される体感がある。（F 社）

・ Google などの AI 開発者は規制に動いていると認識しているがあっているか。（インタビュアー）

→ 大手の AI 開発者に関しては、ルール作りは進めており、防げている印象がある。我々が調査した対象のサービスは、わいせつ画像生成に特化したものがほとんどであり、それらについては、悪化が進んでいると認識している。（F 社）

④ その他

・ 上記以外に、生成 AI 由来のわいせつ物に関する重要な論点などあればご教示いただきたい。（インタビュアー）

→ 生成されたものと本物の境界が曖昧になってきていると感じる。実際に被害児童がいると思われる画像に対して、推測となるが摘発されにくくするために AI 加工のマークをつけたであろう画像を発見したこともある。また、CSAM を取引するダークウェブや匿名性の高い SNS コミュニティは参加するための通行手形として CSAM の提出・提供を求められるケースが珍しくない。さらに、ウォーターマークを付与する場合、AI でウォーターマークを削除する技術も既に存在している。またメタデータに AI 生成の履歴を残すケースの場合、スクリーンショットで撮影すればメタデータが残らないため、有効性には疑問を感じている。（F 社）

G 社（金融サービス業）

① 生成 AI 導入に関する状況

・ どのような業務に対して、どの程度生成 AI を活用しているか。（インタビュアー）

→ 社内では、一般業務（メール作成、翻訳、要約、WEB 検索等）や RAG 検索、代理店では一般業務＋RAG によるマニュアル、約款、商品情報等の回答生成といった業務で利用している。（G 社）

→ 代理店様用の AI システムは主に営業業務で使用されるのか。（インタビュアー）

→ 代理店の業務には、営業だけでなく契約の保全も存在する。お客様からの問い合わせを代理店が受ける場合があるため、それらに回答したり、お礼状のメール文面作成、営業話法の練習にも利用したりしている。（G 社）

→ 代理店様、社内の従業員様は、生成 AI を活用し様々な業務を進められているという状況と理解した。（インタビュアー）

・ 生成 AI を利用したアバターによりお客様向けの問い合わせ回答を行うシステムを先行リリースするとお伺いしている。リリースにあたって課題があればお伺いしたい。（インタビュアー）

→ 社内向けシステムとお客様向け（社外向け）システムでは大きな違いがあると考えている。前提として、今年の初めに AI のガバナンスを強化するために AI リスクの管理規程、要領などを策定し、生成 AI を含む AI 案件については、AI リスク管理部会で審議されることとなった。リスクを 4 段階に分け、どの AI がどの会議体で審議されるかの整備を行っている。当該アバター案件は、一番レベルの高いリスク分類としているため、各 AI リスクにどのような低減策を取るかを事前に検討している。また、金融庁とどのような業務であれば人間（社員や代理店）を介さないでアバターとお客様が直接やり取りできるかの範囲について議論しながら策定している。課題としては、まずハルシネーションの問題がある。対策については、ハルシネーションが起きて、お客様に誤認を与えるような場合、弊社から訂正連絡を実施する運用を検討している。そのために毎日ログをモニタリングすることが必要になり、業務負荷が高いことが課題となっている。（G 社）
→ 承知した。ハルシネーション問題について、お客様向けのシステムの場合、何パーセント程度の精度を担保したい等はあるか。（インタビュアー）
→ 先行事例がないため、ベンチマークとなる数字がないが、現状では、誤回答率は 3 % 未満となっている。（G 社）
→ 本システムについて、社員や代理店様からどのようなお話が上がっているか。（インタビュアー）
→ アバターを介してできる業務の範囲が狭いため、答えられる範囲が限定的である。そのため、回答可能な範囲の拡大要望や、アバターの容貌について選択したいなどの声がある。（G 社）

・ 生成 AI を活用することが効果的であると考えられているコールセンター業務や保全業務について、計画段階か、それとも試行段階か。（インタビュアー）
→ 計画段階のものや、運用段階のものもある。最も効果的であると考えているのは、コールセンター業務であると考えている。コールセンターでは、一般的なお問い合わせのほか、各保全業務の受付など多岐にわたる業務を行っている。また、代理店においては、現在は RAG を使い問い合わせに対応するための生成 AI を活用しているが、今後は提案をサポートする AI エージェントも考えている。（G 社）
→ コールセンター業務は、実際に実証されているフェーズという理解で良いか。（インタビュアー）
→ その通りである。（G 社）

・ 生成 AI における個人情報の取り扱いについて、クラウドベンダーとの契約で担保している理解で良いか。（インタビュアー）
→ その通りである。例えばある AI システムでは、クラウドベンダーと契約を締結した

上でセンシティブ情報の入力は禁止としてガードレールをかけている。その他扱う情報によって、法務相談の上で個別判断している。この辺りは個人情報保護委員会の見解等を踏まえ、適時適切に柔軟に対応していく。（G 社）

・　逆に生成 AI を適用しづらい業務はあるか。（インタビュアー）
→　今のところはない。法務、監査など 2 線業務は他の業務と比べてよく使用している。ただ、現在は弊社では著作権法の侵害に考慮して、画像生成については禁止している。（G 社）

・　生成 AI の利活用状況について、社内累計利用率（1 回以上 AI を利用した社員の割合）は 80%以上である。アンケートベースではあるものの効率化は 30%程度できているという声をいただいている。また、代理店の利用においては、70%以上の代理店から継続して利用したいとの声を得ている。（G 社）

・　生成 AI を使用する際、テキスト生成で使用することが大半か。（インタビュアー）
→　テキスト生成がメインである。著作権の問題があるため、現状は社内で画像生成は禁止している。今後ルールが変わる可能性はあるので、柔軟に対応していく。（G 社）

②　労働環境への具体的な影響
・　生成 AI を活用することによって、各現場における業務の進め方にどのような変化があるか。（インタビュアー）
→　社内での利用率は高く、仕事の進め方として先輩社員に聞いたり、WEB 検索したりするよりも先に生成 AI に聞くという社員が増えている。またアイデアの壁打ち等は AI でできるため、資料作成の効率化と、品質向上に変化がある。
→　生成 AI を活用することで、コミュニケーションが希薄化するのではないかといった懸念があると考えるが、社内で対策等は実施しているか。（インタビュアー）
→　希薄化するとは考えていない。マニュアルを見ずに何でも先輩社員に聞く、ということは減ると考えているが、コミュニケーションの希薄化とは別の問題である。（G 社）

・　生成 AI の活用によって、労働時間・業務量・担当者の心理的負荷など、負荷という面で働き方に影響はあるか。（インタビュアー）
→　業務効率化が進む、あるいは 0 → 1 でのアイデア創出等の心理的負荷は下がると考えられる。一方で、ハルシネーションなどのリスクに対する対応（検証や監視）が必要になってくる。（G 社）
→　ハルシネーションへの対応は、これから本格的に対応される状況か。（インタビュア

ー）

→ すでに検討し、実施している。上述の通り、リスク管理態勢は出来ており、それに対して今後導入を検討している AI や AI エージェントに検証や監視の役割を持たせることが必要になってくると考える。（G 社）

→ ハルシネーションリスクに対応するために、一般業務で AI を用いて壁打ちや要約を行い作成した成果物に対して、従来と異なるレビュー観点を加えるリスク対応を組み込まれているか。（インタビュアー）

→ 新たに対策を追加したということはない。上述の通り、AI のガバナンスを強化するために AI リスクの管理規程、要領などを策定し、生成 AI を含む AI 案件については、AI リスク管理部会で審議しているが、一般業務において AI を用いて作成された成果物に対する追加のレビュー観点は規定していない。（G 社）

・ 生成 AI の活用によって、今後働き方にどのような変化が生じると想定されるか。（インタビュアー）

→ AI エージェントの活用が進むと、AI の検証、監視といったマネジメントや、活用方法を考える業務にシフトすると考えられる。（G 社）

→ AI エージェントの活用について、どこでの活用を考えられているのか。（インタビュアー）

→ 様々な場面での活用を考えている。通常の業務、金融サービスに関する業務、ヒトが実施していた業務については、AI エージェントにすることができると考えている。ただ、お客様の心情に寄りそう業務などは、人間の仕事として残るものであると思う。（G 社）

③ 人材・スキルに関する今後の見通し

・ 生成 AI 導入にあたり、社員への教育（リスキリング含む）や再配置は行われているか、あるいは今後検討しているか。（インタビュアー）

→ 生成 AI に関する研修は月 1 回実施している。再配置については、人事領域でもあるため回答を控えさせていただく。（G 社）

→ 月 1 回の研修について、部署横断で使い方を共有するような場になっているのか。（インタビュアー）

→ 月 1 回の研修は、オンライン講習で交流の場ではない。これとは別に部署毎に特化したワークショップも実施している。また、当部で生成 AI の利活用事例を収集し、社内の掲示板で事例紹介を行っている。（G 社）

・ 人材配置について、貴社内の人員配置はどのように変化していくと想定されるか。（インタビュアー）

→ 個人的な想定であるが、今後は労働集約型業務からよりクリエイティブで戦略的な業務にシフトすると想定している。また、上述の通り、人事領域については、当部から会社としての回答はしかねる。（G 社）

・ 人材活用方針の検討について、AI への対策に人的リソースを割いていく想定で合っているか。（インタビュアー）

→ 世の中の AI リテラシーが上がってくれば、ハルシネーションの許容度が変わってくると思われる。世の中の動向に合わせて柔軟に対応を変えていく必要があると考えている。今は初期の段階で抵抗があるお客様も一定いらっしゃると思われる。お客様を第一に考えたとき、時代時代ですべきことは変わってくる。我々は業界でも先行して最先端の技術を取り入れているが、攻めと守りのバランスをとっていくことが大切と考えている。（G 社）

H 社（IT サービス業）

① 生成 AI 導入に関する状況

・ どのような業務（エンジニア、セールス＆マーケ含む）に対して、どの程度生成 AI を活用しているか。（インタビュアー）

→ 業種によって違いがあるが、エンジニア組織は外向けの記事に出ているように、基本的には AI を活用して高い生産性を出すことを基本的な期待に含めている。セールス&マーケや他ビジネス組織でも AI を使っていない部署はおそらくなく、何かしらの形で活用はしている。エンジニアがコーディングする際、いくつものコーディング補助用の AI ツールを導入しているので、AI の活用を前提にどの AI ツールを使うか、利用者の経験や費用面等を考慮しながら活用している。（H 社）

→ コーディングスキルの高い技術者と初学者では活用度合に違いはあるか。（インタビュアー）

→ 新卒のエンジニア見習いの社員は、AI の技術について勉強したほうが良いなどの話題が上がってくることはないため、特に差はないと考える。（H 社）

・ 他の部門ではどのように AI を活用しているか。（インタビュアー）

→ 商談の書き起こしやお客様の情報を調べるなど、セールス・マーケティングツールに付属している AI ツールを業務に活用している。（H 社）

→ AI を活用することで、手作業の業務は減ってきていると思われる。（H 社）

・ AI ではなく、人が行わなければいけない業務にどのようなものがあると思われるか。（インタビュアー）

→ 現状は、業務の中で AI に代替できそうなものから AI に置き換えている段階である。業務が AI に置き換わった後に、残っている業務を人が行ったほうが良いかどうかを試行錯誤することになると思われる。残った業務が AI に置き換えにくいかどうかは、今のところまだわからない。効率が良いから置き換えるという以上に、多くの議論を経て置き換わっていく可能性はある印象を持っている。(H 社)

・ 現状 AI は、コーディング、事務や調査の業務に活用されて理解で良いか。(インタビュアー)
→ そのような業務に活用しやすい。その他には、定型の質問対応への活用が挙げられる。チャットボットで月に数千件の質問に回答してもらっている。(H 社)
→ 貴社の AI システムの優れているポイントは利用者からのフィードバックを得て改善されていく点か。(インタビュアー)
→ あっている。間違っていると永遠に間違ったままなので、AI をチューニングしていくことが肝となる。(H 社)
→ バックグラウンドにあるナレッジを正しく最新のものにすること、検索に AI を活用することで必要な情報を効率的に取得できるようにする技術を伸ばすことの 2 軸に力を入れてやってきている。(H 社)
→ データの整備は、人手を使って行っているのか。(インタビュアー)
→ データの元を作るのは人が行っており、そこに時間をかけている。(H 社)

・ 貴社が公開されているリリース以外で、生成 AI を活用することが効果的であると考えられる業務があればお聞かせいただきたい。(インタビュアー)
→ 日々AI がものすごいスピードで進化していくなかで、今の状態で「この業務は向いている、向いていない」を判断するのは時期尚早だと感じている。基本的には、どんな業務でも AI を活用したら効果があるかもしれない、という発想を持つこと自体が大切だと考えている。(H 社)

② 労働環境への具体的な影響
・ 生成 AI を活用することによって、各現場（エンジニア、セールス＆マーケ含む）における業務の進め方にどのような変化があるか。(インタビュアー)
→ 現在の業務が生成 AI 活用で、より簡単になる、ミスがなくなる、効率化される、というのはもちろんだと思うが、それ以上の変化、特に今までできなかったことができるようになることを期待している。(H 社)
→ 業務に AI を活用することが前提となっているが、AI を活用することで従来の仕事のやり方と変わって来たところはあるか。(インタビュアー)
→ 変化の具体例としては、プレゼンに向けた練習をする際に AI と一緒に練習を行って

いる中で、AI に資料を読ませることで的確なアドバイスをくれたり、人をやる気にさせたりしてくれる。このような情緒的な価値も感じている。（H 社）

→ 社内で AI を使う前までは、チームのメンバーに声をかけていた。その場合、伝えたいことの背景についてのインプットから始める必要があり、双方で時間を取られてしまうことがコストになっていた。また、自分よりもレベルの高い人をアドバイザーとして呼べるかというと必ずしもそのような状況ばかりではなく、良いフィードバックをもらえるとは限らない。AI であれば、レベルの高いフィードバックがすぐにもらえるため、社内のメンバーだけでやっていた時とは大きく違っている。効率化され、生産性のアップに直結していると考える。（H 社）

→ 効率化だけでなく、質も上がり、更にモチベーションを上げてくれ、仕事を楽しくしてくれている。（H 社）

→ その他では、相手が AI なのでミスが怖くない。当社はミスについてポジティブにとらえるカルチャーなので、率直にフィードバックをもらえるが、それとは別に第三の頼りになる味方ができたという実感がある。（H 社）

→ AI 活用についての ROI を測っていく中で業務効率化やコストなどの観点はあるが、それ以外の上述した情緒的な価値もあるのが AI の特徴と言える。議事録も AI に取ってもらっている。議事録を取ることを意識する必要がなく、議論に集中できる。新人社員が議事録を取ることはなくなっている。（H 社）

→ 会議前に準備する際、自分一人で壁打ちができ、AI から指摘をもらいそれを反映した状態で会議に参加するので、レベルがあがった状態でメンバーと話をすることができ、会議の成果が変わってくると考える。（H 社）

・ 一方で、AI を活用することについて、ネガティブな声があればお伺いしたい。（インタビュアー）

→ AI にどのような情報を与えて良いのか、会社の視点からは様々なリスクが考えられ、一定のガイドラインを作成する必要があると思われる。（H 社）

→ 組織的なところのほか、活用していく上での課題はあるか。（インタビュアー）

→ 課題は多い。AI のリスキリングは当たり前のようにある。勉強会はたくさんあり、そのハードルを下げる啓蒙活動を行う必要がある。（H 社）

・ 労働の代替に関する文献もあるが、代替されることによる従業員目線でのリスクに対する意見があれば、お伺いしたい。（インタビュアー）

→ 当社は、IT の成長企業で毎年新しい仕事が出てきており、また社員もどんどん増えてきている。そのような状況なので自分の仕事が AI に代替されたら、別の仕事ができるようになるという発想。よって、AI を入れたくないと言う社員はおらず、AI に自分の仕事が代替されることへの危機感は起きにくい組織である。会社の規模や状況に

よって感覚は異なると思われる。(H 社)

・ 生成 AI の活用によって、労働時間・業務量・担当者の心理的側面のような観点で、負荷という面で働き方に影響はあるか。負荷の面がない場合、プラスの面でどのような影響があるか。また、現在はまだ変化が確認できていない場合でも、今後どのような変化が生じると想定されるか。(インタビュアー)

→ 業務のあり方、やり方が根本的に見直されると思うので、今までと同じやり方ではなく新しいやり方にしていくという変化が起きるのは間違いなく、それが負荷になるか、プラスになるかは進め方や捉え方次第ではないか。(H 社)

③ 人材・スキルに関する今後の見通し

・ 生成 AI 導入にあたり、社員への教育(リスキリング含む)や再配置は行われているか、あるいは今後検討しているか。(インタビュアー)

→ 社内への啓蒙活動は必要だと思っている、今後力をいれていく想定である。(H 社)

・ 生成 AI 活用に関連して、貴社内の人員配置はどのように変化していくと想定されるか。(インタビュアー)

→ 今の業務ベースの人員配置の考え方ではなく、AI と共に働くことを前提にした、業務、組織設計が必要になってくると考えている。(H 社)

・ 社員の配置転換について、検討している、もしくは今後検討する必要が出てくるとの考えはあるか。(インタビュアー)

→ AI が出てきたので会社の組織を一律に変更することは考えていない。現場で AI を活用することでより少ない人員で業務を遂行できるようになった場合、社員の配置転換もあり得る。(H 社)

・ AI を活用することについては、部署ごとに任せているのか。それとも、AI 活用を推進していく組織があるのか。(インタビュアー)

→ エンジニアの部署は、その組織自体で積極的に AI の活用を推進している。また、会社としても全社の AI スキルを上げていくために推進している。(H 社)

→ エンジニアの部署のように AI 活用に積極的に取り組んでいる組織もあれば、そうでない部署もある。特に後者の組織の AI スキル向上に注力している。(H 社)

→ AI 活用に積極的に取り組んでいる組織があるということだが、統制についての課題はあるか。(インタビュアー)

→ 会社としての一定の承認フローがあるが、AI 活用を早く進めるため、一部の特にリテラシーの高い社員に対しては承認の基準を下げ対応している。(H 社)

・　生成 AI 活用による人材活用方針の検討について、貴社内では期待・懸念の両面でどのような声が聞かれるか。（インタビュアー）

→　今後社内の声を集めていく想定である。期待の面では、AI 活用についての期待が大きい。（H 社）

I 社（ファクトチェックの非営利団体）

①　AI による偽誤情報の実態と傾向

・　貴組織に寄せられる偽誤情報のうち、偽情報（意図的に作成された間違った情報）、誤情報（意図的ではないが間違っている情報）の割合はどの程度なのか。（インタビュアー）

→　我々が検証する偽・誤情報の大半は自分たちで見つけるもの。偽情報と誤情報の割合は分からない。なぜなら「意図」があるかどうかは情報を作成した本人しかわからないから。（I 社）

・　貴組織に寄せられる偽誤情報はどの程度あり、そのうち AI 由来のものの割合はどの程度であるか。（インタビュアー）

→　検証した事例のうち、どれだけが AI 由来のものかは分からない。テキストのものは AI かどうかがわからないため。画像や動画で言えば、まだチープフェイクの方が多いが、2024 年後半から AI 由来が増えていると感じている。星の数ある偽情報・誤情報を全部検証するのは不可能。割合は分からないとしか言えない。日本は他国と比べてペースが遅い。2024 年に AI による偽情報が氾濫しディープフェイク元年になると言われていた。しかしチープフェイクの方が多かった。人が見たくなるようなものを作る能力はまだ人間の方が高い。（I 社）

→　SNS において、AI で作成した災害情報の偽画像で混乱を招いた事例があると認識している。たまたま AI で生成したものが拡散したのか。どういった見解があるか（インタビュアー）

→　デマが広がるときの公式があり、それは状況の注目度と不確かさである。災害の時は偽情報を流す人たちにとってチャンスであると考えている。（I 社）

・　AI が普及する前後、また AI が普及し始めて以降で偽誤情報のチェック依頼や通報件数はどのような遷移になっているのか。また質的な変化はどのようなものか。（インタビュアー）

→　メール、フォーム、LINE での問い合わせが来る。偽情報が増えているというより、

我々の知名度が上がって問い合わせが増えていると考えられる。我々が検証している数に関しては、2022 年は月 10 本で、2025 年は月 40 本。偽情報が増えたより、処理能力が上がった。AI 由来のものは増えており、また質も高まっているが、まだ人間が見分けられるものの方が多い。AI による判定ツールの信頼性の方が人間よりも低い。（I 社）

→ リテラシーを高めていくことが重要なのか（インタビュアー）

→ ファクトチェックは対策として不十分。ツール開発、メディアリテラシー教育も必要。全部対策を行っても解決しない。状況がこれ以上悲惨になることのスピードを落とすぐらいの効果であると考える。(I 社)

→ AI による判定ツールはなかなか精度が出ないのか（インタビュアー）

→ ツールによって判定が異なる。また人間の顔には強いが、風景には弱いこともある。まだ限界がある。また、65%ディープフェイクと言われて、ツール利用者はどういう判断して良いか分からず、ファクトチェックに使えない。(I 社)

・ 最近、AI によって作成された可能性が高い偽誤情報で、特に注目された事例はあるか。(インタビュアー)

→ 特に注目したものはないが、海外の災害情報や政治情報は AI 由来が特に増えている。SNS 上で日本語と英語で同じ設定でディープフェイクの傾向を見ている。海外は政治ものが多い。日本は同じキーワードでもわいせつ物系の画像が多い。国によってどういう偽情報が広がりやすいのか、特徴が違う。インドは音声のディープフェイクが多い。アメリカは動画のディープフェイクが多い。(I 社)

・ AI による偽誤情報の主要な流通経路として顕著なものがあればご教示いただきたい。（インタビュアー）

→ 偽・誤情報の主な流通経路は X から YouTube と TikTok に変化している。偽誤情報は国ごとに傾向が違うため、ここで個別に言及することは不可能だが、日本は AI 由来のものがまだまだ少ない国だと感じている（I 社）

→ 世に使われているプラットフォームの動きと追従して、偽情報の拡散の経路が変わると理解したが、合っているか（インタビュアー）

→ あっている。ただ、コンテンツのフォーマットは現状テキストか画像が音声か動画である。将来は AR (Augmented Reality)、VR (Virtual Reality)も入ってくると想定されるが、コストパフォーマンスが悪い。現状動画で偽情報を拡散するのが合理的であると考える。(I 社)

② AI による偽誤情報への対策

・ 偽誤情報か否かを識別するためにどのような手法をとっているのか。（インタビュアー）

→ 関連情報との比較、AI による描写ミスのチェックなど。AI 判別ツールも使うが、参考程度である。（I 社）

→ ディープフェイクは色々あるが、音声は見分けるのは難しいなど、他に見分けるものが難しいとかあれば、お聞きしたい（インタビュアー）

→ 音声は難しい。悪魔の証明になる。言ってないということも証明できない。音声の出所が分からないぐらいしか言えないことが多い。（I 社）

→ 他の種類のディープフェイクに関しては、どうなのか（インタビュアー）

→ 現状は多くのものは不自然なものが多い。ディープフェイクの検証はやりやすい。動画の場合、音声だけでなく絵もあるからヒントが多い。繰り返しになるが、これから大変になる。無尽蔵に作れるようになるため、手に負えなくなるというのは見えている未来である。AI による判定、または取り締まる法令がないと、どうしようもなくなってしまう。（I 社）

→ AI による判定、法令での取り締まり、リテラシー教育、研究、など多様な対策に関して、濃淡をつけるなら、どのあたりが重要になってくるのか（インタビュアー）

→ 全部重要になる。各関係するアクターが一生懸命やるしかない。比較的にやらなくて良い対策というものはない。（I 社）

→ 現行法令でどういった課題感があるのか。法観点で何か考えがあれば、お聞きしたい（インタビュアー）

→ 言論の自由があるため議論していくことが重要。実効性の高いルールがないまま進むと大変になる。そのため、何かしらのルールが必要。どこまで厳格なものを作るのか、議論しないと実効性がない。（I 社）

・ AI による偽誤情報を検出する際の課題はあるか。（インタビュアー）

→ 偽誤情報は無数にあるため、AI 由来に限らず、検出は困難である。特に動画由来のものは動画内のどの部分が検証対象になりうるかを判断するのに時間がかかる。X が問題の中心の時は文字数制限があり、簡単だった。YouTube は 30 分動画。文字起こししたら、文字数が長い。（I 社）

→ 現状、どれぐらいの時間が掛かるのか（インタビュアー）

→ これを検証した方が良いと判断するのに数時間かかる。その間に百万再生されてしまう。また、検証時間は 30 分で終わるものもあれば、取材などを行い、数週間かかるものもある。記事に起こす場合、さらに時間がかかる（I 社）

・ ファクトチェックの自動化や半自動化に AI を活用している、または検討している事例はあるか。（インタビュアー）

→ 「ファクトチェックの自動化」は現状の技術では不可能だと考えている。あるテキストや画像や動画が AI 由来であることを検出できたとしても、その情報が「誤っている」かどうかとは別であるため。しかも、AI 由来かどうかの判定も現状では人間よりも劣る事例が多い。画像や動画ではない、テキストの論理の判断などに関しては、AI による判定は参考にはなるが、最終的な判定にはほど遠い。仕事でも AI を使っている。ただ、判定はできない。判定するのは人間。AI は下調べで、最終判断は人間が行う。（I 社）

・ 一般人が使えるツールとして、X での Grok があるが、一般人は何が利用できるのか。プラットフォーマーが実装すべきなのか。（インタビュアー）

→ Grok は使えるが、よく間違える。第一にやることはメディアリテラシー。AI は優秀で 24 時間働くが、詰めが甘い人とよく説明している。そのため、仕事は振るが、最後は自身で確認する。重要なポイントは、ⅰ）発信源、ⅱ）情報の根拠、ⅲ）関連情報、の 3 点を調べることである。（I 社）

・ その他、AI 由来の偽誤情報の現状把握や対策において、企業や一般の読者が認識すべき重要だと考えられる論点があればご教示いただきたい。（インタビュアー）

→ 現状でディープフェイクの割合がそれほど多くないとしても、今から準備しなければ間に合わない。今後、指数関数的に増えていくと考えられる。一人一人がメディアリテラシーを身に着ける必要があり、ツール開発、ルール整備を実施していかないと状況が悪くなる。2016 年からファクトチェック・メディアリテラシーの活動が本格化したが、状況が悪くなっている。対策の進化より状況の悪化の方が早い。（I 社）

③ AI による偽誤情報に対する社会的役割・啓発

・ AI によって「真実らしさ」が増した偽誤情報に対し、一般の読者が持つべきリテラシーとは何だと考えるか。（インタビュアー）

→ あと数ヶ月、数年のうちには「人間の目で確認する」という我々の知見では対応できないディープフェイクが増えると考える。重要なのは、「画像や動画や音声があったとしても、それが事実とは限らない」ということを前提とすること、発信源や関連情報を確認するという基本動作を身につけること。とにかく世の中には大量の間違った情報が拡散している。研究結果によると、人々は偽誤情報の内、14.5%しか間違って

いると認識できないとされている。(I 社)

・ AI による偽誤情報に関して今後どのような啓発活動や提言を重視していきたいか。(インタビュアー)

→ 日々のファクトチェックを通じて、どのようなディープフェイクが増えているかを明らかにし、実践的なメディア情報リテラシー教育にもつなげるとともに、ツール開発や法的規制の議論にも貢献していく計画である。(I 社)

→ ツール開発においては、どのように関与されているのか。(インタビュアー)

→ βテスターとして、ツール開発に貢献している。(I 社)

J 社（報道機関）

① AI による偽誤情報の実態と傾向

・ 貴社で作成したことを名乗る偽情報がネット上で流通したことがあるか。ある場合、AI によって作成された偽情報であることが把握できた件数はどの程度あるか。(インタビュアー)

→ 当社の社員でない者が、ニュースサイトのロゴマークや著名人の氏名を無断で使用し、あたかも著名人が投資を勧めているような内容の記事であるかのようにユーザーを誤認させて投資などを誘うサイトがあった。「購読料金の請求を装った不正メールにご注意ください」という注意喚起をリリースしたこともあるが、すべてを把握できているわけではない。また、記事の無償版のところだけをインプットに AI が記事を作成し、有償版の部分を無視した状態で出力されるような事例が確認できている。(J 社)

→ AI により生成された偽情報の件数が急増して困っていることはないか。(インタビュアー)

→ 会社として、AI により生成された偽情報の件数を統計的に把握したりしているか承知していないため、分かりかねる。(J 社)

・ 貴社が報道機関としての業務を行う場合にネット上で情報収集する際、入手情報が偽情報/誤情報であったことはどの程度か。その情報が AI による偽情報/誤情報であることが把握できたことはどの程度あるか。(インタビュアー)

→ 当社は、多くの従業員が従事しており、情報収集においてどの程度偽情報が含まれているか、日々一人ひとりの情報収集の活動について調査を行っていない。

そもそも取材活動は、一次情報にあたることが原則であるため、SNS が普及した後でも、普及する前からの基本動作は変わらない。偽情報が増えて取材活動に著しい影響が出ているということはないと思う。

ニュースを報道する上では、事実関係を確認しなければならない。事実関係を確認するのは、一次情報である。SNS 上で情報収集を行い、参考にすることはあるが、決して鵜呑みにすることはない。記事を書くとき、X で検索してそのまま記事にすることはない。公的なもの、発表されているもの、過去のデータベースなど、信頼性があり根拠のあるものを元に記事を書くことが大前提となる。

偽情報が増えていることは、その通りであるが、記事を書くにあたり阻害されているということはないと思う。

ただ、報道内容などが間違って解釈され、拡散されるなど、偽誤情報が SNS で出回るケースが増えており、ソーシャル上の投稿をチェックする手間も増えている。

また、社会に著しく影響を与えている偽情報があれば、ファクトチェックを行いこれは誤りであることを我々の取材リソースを活用して多くの方にお知らせする活動を行っている。(J 社)

→ 提供された情報をそのまま記事にすることは絶対にあり得ない。その人がそこに本当にいるのか、その時に撮った写真なのか、必ずその人にアクセスして信頼性が高いものなのかどうかを確認することが一番大切なところ。写真の加工が見抜きにくくなるであろうという懸念は持っているため、より慎重に確認していかなければいけないという話はしている。

報道機関としてどれが正しいのか情報のチェックをしてもらいたいという世の中からの要請があり、報道の役割として以前は行っていなかったファクトチェックにもより力を入れて行うようになった。これは嘘ではないかとファクトチェックしたところ、本当であったということはあった。(J 社)

→ 同業他社でなりすましのサイトを本物と誤認して、そのまま記事にして誤った報道をしてしまったという事例はあったが、それは SNS による偽情報が増えたからではなく、新聞社の報道機関としての基本動作での裏取りが不十分であったことが原因ではないか。(J 社)

・ 実際に、AI による偽誤情報に触れた、騙された、検証に時間を取られた経験はあるか。(インタビュアー)

→ どのような情報でも同じように一次情報に当たるという対応をしているため、変化はないと思う。(J 社)

・ AI で作られたフェイク音声・画像・動画など、視覚・聴覚情報の信頼性に対する懸念は現場で高まっているか。(インタビュアー)

→ ディープフェイクは、一般の方、記者も含めて見抜きにくいため、警戒しなければならない。ディープフェイクの画像、音声の精度が高まっているため、気をつけなければいけないという認識は共有されている。(J 社)

② AI による偽誤情報への対策

・ 貴社では AI で作成された偽情報にどう対処しているか。偽誤情報の入手（偽誤情報を掴まされる）、偽情報の拡散（貴社を名乗る偽記事の拡散）の二つの観点で対処している事項をご教示いただきたい。（インタビュアー）

→ 偽誤情報を入手した場合は、一次情報の確認を行う、偽情報の拡散は公式サイトから注意喚起を行うことが重要ではないか。（J 社）

・ 報道のスピードと正確性のバランスが難しい中で、AI による偽誤情報に「一度乗ってしまう」リスクへの対策として、どのような活動を実施しているか。もしくは、どのような活動が必要と考えるか。（インタビュアー）

→ スピードを優先するために正確性をないがしろにすることはない。事実関係が確認できた内容を報道することが大原則である。すべてにおいて正確性を優先している。（J 社）

→ 政府、AI 提供者、開発者、一般市民といった様々な目線でできる対策について、コメントをいただきたい。（インタビュアー）

→ 個人的な意見として簡単に申し上げれば、SNS のプラットフォーマーが責任や役割を果たすべきであると考えている。一方で、偽情報の投稿を削除したり、アカウントを削除したりすることは表現の自由を損なうことになりかねないため、バランスを取ることが難しい。政府が介入することは、表現の自由を損なう懸念がある。信頼できるメディアが発信している情報を目立たせ、一般の読者にはそれを理解してもらうことが重要だと考える。

　総務省の有識者会議でも偽情報は新しい情報よりも何倍も拡散が速いという指摘が出ている。読者の皆様には、リテラシーを身に着けてもらうことが一般論として重要と考えている。また、オリジネータープロファイルというインターネット上の情報（コンテンツ）の作成者・発信者を、ユーザーが確認するためのデジタル技術を用いてメディアに認証を付与する取り組みを行っている。（J 社）

・ 一般市民に対しては、信頼性ある新聞社の記事を確認するよう啓発していく必要があるという理解であっているか。（インタビュアー）

→ 個人的な意見となるが、我々が考えるニュースは紙の新聞やニュースサイトであったが、今の若い世代は Instagram、TikTok から情報を収集したり、記者やジャーナリストよりもインフルエンサーが影響力を持っていたりする印象がある。中高年も YouTube を情報源にしていることが多いようだ。新聞社の信頼できる記事に触れていただく必要性は高まっていると考えている。（J 社）

・　記者個人が AI を使う際に、社内で定めているガイドラインやポリシーはあるか。ある場合、どのような観点を重視しているか。参考にした資料や報道機関特有の考慮事項があれば、可能な範囲でご教示いただきたい。（インタビュアー）

→　ガイドラインはある。「利用できる AI のサービスは会社が認めたものだけ」「出力された生成物は必ず事実確認する」「生成物を記事などでそのまま使う場合は断りを入れる」などを定めているが、詳細は社外秘となるため話すことはできない。（J 社）

③　AI による偽誤情報に対する社会的役割・啓発

・　AI が偽情報を量産・拡散する時代において、貴社が果たすべき情報発信の指針としての役割をどう再定義しているのか。（インタビュアー）

→　報道機関は、一次情報にきちんとあたり、確認できた正確なものを報道することを長年積み重ねてきたが、ファクトチェックという取り組みを始めたのは、「この情報は違う」というところまで積極的に踏み込んで発信していくことが読者の期待に応えるものであると考えているからである。（J 社）

・　報道機関として、貴社が他メディア・SNS・ファクトチェック団体・教育機関などと連携したい／している具体的な取り組みは何かあるか。（インタビュアー）

→　情報交換などを行っている。公式でないものも含めてメディア同士で情報交換を行っている。大学など学校からの要請があるので積極的に行っている。（J 社）