生成 AI 利用における３つのリスク

明治大学ビジネス情報倫理研究所客員研究員
（元内閣官房上席サイバーセキュリティ分析官）
守屋　英一

はじめに

従来のArtificial Intelligence（以下、ＡＩ）は、一定のパターンや法則に従って自動化された出力を行うものであった。一方、生成ＡＩは、ユーザーからのリクエストに対して、自然な文章での回答や画像の生成ができる。例えば、マイクロソフトなどが出資するOpenAI社が2022年11月に提供を開始した生成ＡＩのChatGPTが有名である。ChatGPTの登場により、従来手動で行っていた情報収集、情報分析、文章・画像・音声生成といった一連の作業が自動化された事により、人はより創造的な仕事に集中して能力を発揮することが出来るようになった。但し、生成ＡＩを利用する上で注意すべき点がある。本稿では、生成ＡＩを利用する上での３つのリスク「活用する上でのリスク」、「悪用されるリスク」、「規制によるリスク」について解説する。

活用する上でのリスク

帝国データバンクが実施した「生成 AI の活用に関する企業アンケート」によれば、61.6%の企業で生成ＡＩの活用を検討していると回答しているが、業務で活用している企業は、大企業で13.1%、中小企業で8.5%、小規模企業で7.7%と企業における生成ＡＩの活用は、低いと言える（図1）[1]。生成ＡＩの利用において、どのようなリスクが存在するか分からないため、利用を躊躇しているとも考えられる。ここからは、企業が生成AIを活用する上でのリスクについて解説する。

情報漏洩

多くの生成ＡＩでは、入力した質問や情報が履歴として閲覧する事が出来る。例えば、議事録を作成するため、音声データをテキストに変化した情報を生成ＡＩに入力した場合、入力情報は履歴として生成ＡＩのサーバー上に保管される。そんな中、サイバーセキュリティ会社の調査によれば、生成ＡＩのアカウント情報10万件以上がダークサイトで転売されている事が明らかになっている。アカウント情報を用いて生成ＡＩにログインし、履歴を閲覧される恐れがある。情報漏洩を防止するには、定期的なパスワードの変更や履歴情報の削除などの対応が必要である。

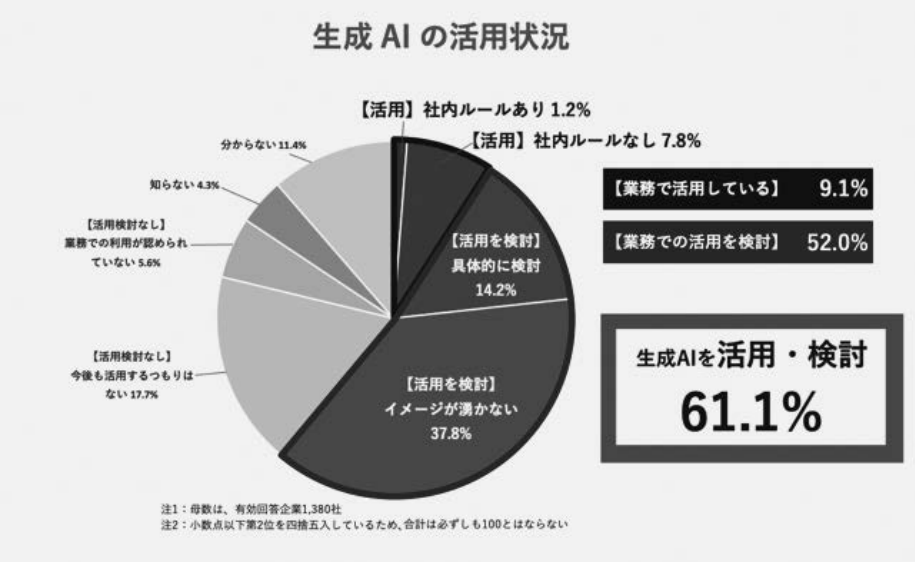

図１　生成ＡＩの活用状況（帝国データバンク調べ）

[1] 生成AIの活用に関する企業アンケート（https://www.tdb.co.jp/report/watching/press/pdf/p230608.pdf）

著作権等の侵害

画像生成ＡＩで生成された画像が既存の著作物と類似している場合は、著作権侵害に抵触する恐れがある。

また、企業が生成ＡＩを利用する際に他人の著作物を含むデータを生成ＡＩに入力する行為において、著作権者の承諾なく行なった場合は、原則として複製権侵害となる恐れがある。例えば、画像生成ＡＩサービスを提供するStability AI社は、機械学習の素材として、デジタル画像を提供するGetty Images社の画像データを無断で使用したとして、2023年2月3日に、Getty Images社から訴訟されている[2]。

ハルシネーション

生成ＡＩは事実に基づかない情報を生成する。この事を生成ＡＩが幻覚（＝ハルシネーション）を見ているかのように、もっともらしい嘘を出力する行為を指す。例えば、生成ＡＩにプログラミング言語「Python」に関する227個の質問のうち、80個以上で実際には存在しないOSSを利用するよう推奨したことが報告されている[3]。

また、初期の生成ＡＩで筆者について質問したところ「日本国内でのテロ行為を計画していたとして、逮捕・起訴され、2009年に懲役20年の判決を受けた」という結果が出力された（図2）。このように情報が必ずしも正しいとは限らないため、その点を理解して利用する必要がある。

悪用されるリスク

次にサイバー犯罪の側面から生成ＡＩの悪用事例として「ディープフェイク」、「マルウェアの作成」、「サイバー犯罪用の生成ＡＩ」について、解説する。

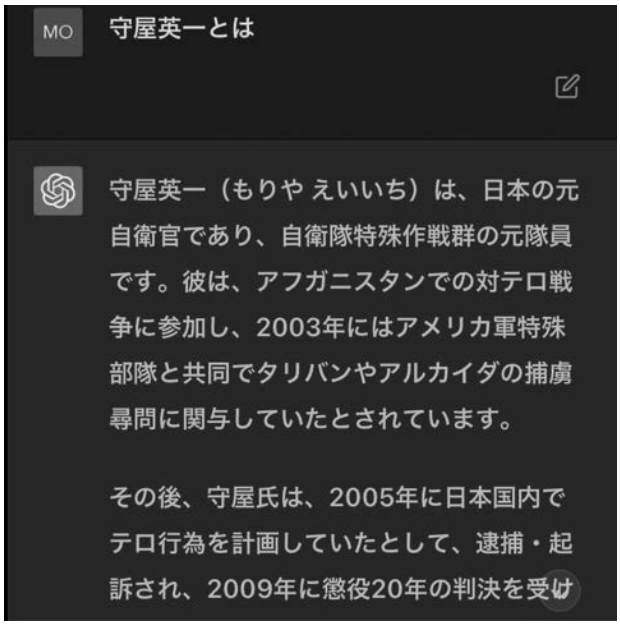

図２　生成ＡＩで筆者について質問した結果

ディープフェイク

海外では、経営者や役員の声を生成ＡＩによって作成して、経理や財務担当者を騙して、偽りの銀行口座に資金を送金させるなどの被害が確認されている[4]。また、2023年4月、米国で誘拐事件が発生。母親は電話越しに、娘の泣き声や叫び声、懇願する声などが聞こえたと証言していたが、実際に娘は誘拐されておらず、娘の声を元にして作成された偽りの音声が用いられたのではないかと言われている[5]。

上記以外にも、生成ＡＩでネガティブキャンペーンやプロパガンダを目的とした動画が目立っている。例えば、昨年の11月に岸田総理の声や映像を使ったフェイク動画がSNSに投稿され話題になった。これは、実在する

[2] 画像生成AI「Stable Diffusion」をGetty Imagesが著作権侵害で提訴、これで２回目の法的手続き（https://gigazine.net/news/20230207-getty-sues-stability-ai/）

[3] 生成AIでソフト開発者攻撃の可能性、架空の返答を悪用（https://www.nikkei.com/article/DGXZQOUC087WL0Y3A600C2000000/）

[4] CEOになりすましたディープフェイクの音声で約2600万円の詐欺被害か（https://japan.zdnet.com/article/35142255/）

[5] I've got your daughter': Mom warns of terrifying AI voice cloning scam that faked kidnapping（https://www.wkyt.com/2023/04/10/ive-got-your-daughter-mom-warns-terrifying-ai-voice-cloning-scam-that-faked-kidnapping/）

人物が発言していない内容をあたかも本当に話しているかのように見せかける行為である。生成ＡＩによって作成された動画は、瞬きが極端に少なかったり、逆に多すぎたりするなどの特徴があるため、騙されないように注意する必要がある[6]。

マルウェアの作成

生成ＡＩは、プログラムコードを生成することが可能である。例えば、「電卓の機能を実装したプログラムを作成してください」と生成ＡＩにリクエストすることでプログラムコードを得ることが出来る。攻撃者は、生成ＡＩを悪用してマルウェアの作成に活用したり、プログラムコードの改善や最適化に用いられたりすると言われている[7]。但し、「マルウェアの機能を実装したプログラムを作成してください」と生成ＡＩにリクエストしても、「法律に違反する行為であり、強く禁止されています」とプログラムコードの作成が拒否される。そのため、機能ごとにプログラムコードを作成し、それぞれのプログラムコードを最後に結合するなどして、禁止行為が回避されている[8]。

サイバー犯罪用の生成ＡＩ

WormGPTとは、マルウェアの作成や、説得力のあるフィッシングメールの作成ができるツールである。ChatGPTでは、サイバー犯罪者がツールを悪用することを防ぐための機能制限がある。但し、WormGPTには、そのような機能制限がないため、サイバー犯罪者の支援ツールとして注目されている。

規制によるリスク

生成ＡＩの急速な普及に伴い、悪用や情報漏洩、著作権侵害、ハルシネーションなどが問題になっている。生成ＡＩによる新たなリスクに対応するため、各国では、システムの透明性、プライバシーの保護、差別の禁止等を踏まえて生成ＡＩへの規制強化が進められている。企業は、生成ＡＩを活用した製品やサービスを提供する上で各国の規制やガイドラインに従わなければ、罰金や訴訟等の問題に発展するため、注視する必要がある。ここからは、欧州、中国、米国における生成ＡＩへの規制の状況について解説する（表1）。

欧州

欧州では、ＡＩ利用時の基本的人権の保障と安全性の確保するため、2023年6月に欧州議会本会議で「AI規則案」が採択された。この法案の特徴は、ＡＩのリスクレベルに応じて規制とAIシステムへの要求事項として「リスク管理システム」「データガバナンス」「技術文書、ログ管理」「透明性」「人による監視」「サイバーセキュリティ」などの遵守が求められている。また、本規則は、EU域外の企業も適用対象であり、規則に違反した場合は、最大3500万ユーロ又は（日本円で約55億円）年間世界売上高の７％の罰金を求めている点が特徴である。

中国

中国では、生成ＡＩを用いた国家の安全や社会秩序を脅かす行為を防止するため、生成AI規制を定めた法律として「生成形人工知能サービス管理暫定弁法」がある[9]。この法律は、生成AIを提供する事業者に対して、アルゴリズムの透明性確保するため届出制度を設けている。また、国家の安全や社会秩序を脅かすような内容を含む場合、当局による事前審査を義務付けている点が特徴的である。

[6] Detect DeepFakes: How to counteract misinformation created by AI（https://www.media.mit.edu/projects/detect-fakes/overview/）

[7] 過度な期待と現実：サイバー犯罪のアンダーグラウンドにおけるChatGPTを中心としたAIの動向（https://www.trendmicro.com/ja_jp/research/23/i/hype-vs-reality-ai-in-the-cybercriminal-underground.html）

[8] I built a Zero Day virus with undetectable exfiltration using only ChatGPT prompts (https://www.forcepoint.com/blog/x-labs/zero-day-exfiltration-using-chatgpt-prompts)

[9] 生成形人工知能サービス管理暫定弁法（https://crdb.jp/wp/wp-content/uploads/2023/07/cdil-15-10.pdf）

表1　生成AIへの規制

	欧　州	中　国	米　国
名　称	AI規則案[10]	生成型人工知能サービス管理暫定弁法[11]	AI権利章典[12]
目　的	基本的人権の保障	安全保障、社会秩序維持	米国人の市民権を保護
特　徴	EU域内にAIシステムを提供する域外企業も適用対象。 違反の場合、最大で3500万ユーロ又は年間世界売上高の7%の罰金。	社会主義の核心的価値観を堅持し、国家転覆の扇動、社会主義体制の転覆、国家の安全と利益の危険、国のイメージの害、分離主義の扇動、国家の統一と社会の安定の侵害、テロリズムまたは過激主義の擁護、民族憎悪、民族差別、暴力、わいせつ、ポルノ、虚偽および有害な情報の擁護など、法律および行政規則で禁止されているコンテンツを作成禁止。	AIを含む自動化システムの設計、使用、導入の指針となるべき5つの原則を特定 ・安全で効果的なシステム ・アルゴリズム由来の差別からの保護 ・データのプライバシー ・ユーザーへの通知と説明 ・人による代替手段、配慮、フォールバック
透明性	システムの設計者、開発者、配備者は、システム全体の機能と自動化が果たす役割、そのようなシステムが使用されていることの通知、システムに責任を持つ個人・組織などを明確に説明する文書を広く一般に提供する。これらの情報は最新の状態に保ち、重要な使用例や主要機能の変更についてはシステムの影響を受ける人々に通知する。	サービスの種類の特性に基づき、生成AIサービスの透明性を高め、生成コンテンツの精度と信頼性を高めるための効果的な対策を講じる。	システムの機能と結果についてタイムリーに説明を提供するべきである。また、システムの機能変更によって影響を受ける人々に通知する必要がある。情報は可能な限り公開されるべきである。
プライバシー	個人の合理的な期待に合致し、厳密に必要なデータのみを収集する。システムの設計者、開発者、配備者は個人からの許可を取得し、データの収集、使用、アクセス、移転、削除に関する個人の決定を尊重する。個人の同意を求める際は、簡潔で、平易な言葉で理解できる内容にする。健康や仕事などに関わる機微なデータについては、より強い保護措置を講じる。	他人の適法な権利と利益を尊重し、他人の心身の健康を危険にさらしてはならず、肖像、名誉、名誉、プライバシー、個人情報など、他人の権利と利益を侵害してはならない。	データは適切に保護され、使用方法はユーザーの同意に基づくべき。データ収集は必要最小限に留め、同意は明確で理解しやすい形で得られるべきで、データの広範な使用についての通知が必要。機密領域のデータは厳重に保護され、その使用は倫理的レビューと制限が必要。
差　別	システムが人種、性別、年齢などに基づいて不当な待遇をもたらすことがないよう、設計者、開発者、配備者はシステムを公平な方法で使用・設計するための継続的な措置を講じる。	サービスの提供の過程において、民族、信条、国、地域、性別、年齢、職業、健康などによる差別を防止するための効果的な対策を講じる。	アルゴリズム差別は、人種、肌の色、民族、性別、宗教、年齢、国籍、障害、退役軍人の地位、遺伝情報などの特定の分類に基づいて不利な影響を与える場合に法的保護に違反する可能性がある。公平なシステムの使用と設計のために、積極的な措置を講じる。

[10] 人工知能に関する統一規則(人工知能法)の制定と特定の連邦立法法の改正
（https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206）

[11] 生成型人工知能サービス管理暫定弁法（http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm）

[12] 12Blueprint for an AI Bill of Rights（https://www.whitehouse.gov/ostp/ai-bill-of-rights/）

米国

ホワイトハウスの科学技術政策局（以下、OSTP ）は、ＡＩの開発に関する原則をまとめた「AI権利章典」を発表。OSTPは、AI技術がイノベーションを促進している一方、人々が自動化されたシステムによって監視や順位付けされることが増えていると指摘。また、多くのアルゴリズムが偏見を持ち差別的なデータ処理を行っていると問題視している。OSTPは、大手テック企業の説明責任を追及し、米国人の市民権を保護する取り組みと位置付けている。

対策

生成ＡＩを利用する上での３つのリスク「活用する上でのリスク」、「悪用されるリスク」、「規制によるリスク」について解説してきた。ここからは、リスクを回避するための対策について解説する。

① 利用規約の確認

生成ＡＩ開発元へ情報流出を防止するには、例えば、ChatGPTの場合は、ＡＰＩ経由でChatGPTの機能を利用することで回避できる。OpenAI社の利用規約では、ChatGPTのAPIにおいては、ユーザーのインプット情報や生成ＡＩによるアウトプット情報は、開発元に提供されないため、情報漏洩の不安を払拭することが出来る。

② アクセスを制限

生成 AIは、企業の重要な情報や個人のデータを扱う可能性がある。第三者からの不正アクセスを防止するため、アクセスを制限し、基本的なセキュリティ対策を行う。

③ 入力情報の削除

履歴から情報漏洩を防止するため、生成ＡＩを利用するためのルールにおいて、入力した情報は定期的に消去する事を定める。具体的には、プログラムによる定期的な削除や従業員へのルール教育を継続的に実施する。

④ 事実の確認（ディープフェイク、ハルシネーションへの対応）

生成ＡＩの嘘を見抜くには、信頼できる公的機関や行政のサイト、専門家が運営しているサイト、企業のサイト、新聞記事、論文や学術記事など複数の情報源から確認する必要がある。また、生成ＡＩによって生成された偽りの映像は、瞬きが極端に少なかったり、逆に多すぎたりするなどの特徴がある。

⑤ 利用ルールの整備と教育

生成ＡＩを利用する上で、他人の著作物、社内情報、営業秘密および個人情報等を生成ＡＩを入力してはならない。このように企業で生成ＡＩを利用する際は、利用ガイドラインを整備して、従業員に対してルールの徹底を図る必要がある。一般社団法人日本ディープラーイニング協会では、生成ＡＩの活用を推進するため、2023年5月1日に利用ガイドラインのひな形を作成、一般に公開している[13]。このガイドラインを基に、自社の利用方法と照らし合わせて、加筆･修正することでルールの整備が行える。

まとめ

OpenAI社が提供する生成AIのChatGPTは、従来の自動応答を行うプログラムとは異なり、ユーザーが投げかけた質問に対して、人と会話するように回答が生成される。そんな中、生成AIを活用した生産性の向上が人手不足緩和の一助になると期待されている。しかし、本稿では、生成ＡＩにおける新たなリスクについて解説した。リスクへの対応として、生成ＡＩへのアクセス制限や不正アクセスが発生した場合に備えて、普段から入力した情報を削除するなど、生成ＡＩを利用する上でのガイドラインを整備し、リスクを回避しながら生産性の向上に努めて頂きたい。また、生成ＡＩを用いた製品やサービスの開発を進めている企業もあると思うが、各国において生成ＡＩへの規制が強化されているため、違反による罰金や訴訟等の問題に発展しないように法規動向にも注視して頂きたい。

[13] 生成AIの利用ガイドライン（https://www.jdla.org/document/#ai-guideline)