会誌「情報処理」Vol.66 No.2（Feb. 2025）「デジタルプラクティスコーナー」

特集号招待論文

# 生成AIのセキュリティリスクと研究動向

染谷実奈美[1]　菅　和聖[2]　大塚　玲[1]

[1]情報セキュリティ大学院大学　[2]日本銀行金融研究所

ChatGPTをはじめとする生成AIは，かつてないスピードで社会に浸透しつつあるが，それと同時に，さまざまなセキュリティリスクが社会で顕在化し，研究コミュニティにおいても関心を集めている．これには，誤回答・誤判断のリスクや，プライバシー情報や機密情報の漏洩および差別的表現の出力などの倫理的問題，さらには偽・誤情報の拡散や詐欺などのサイバー犯罪への悪用などがある．生成AIが社会に広く受け入れられていくためには，これらの問題を理解し，適切に対処していく必要がある．本稿では，生成AIがもたらすセキュリティリスクと研究動向について述べる．なお，本稿の内容は，筆者らの所属組織の見解を示すものではない．

## 1. 生成AIとセキュリティリスク

近年，人工知能技術が急速に発展し，文章や動画像などのコンテンツを作成する生成AIが注目を集めている．生成AIの技術的本質は深層学習モデルである．出力データの形式が単純なクラス分類問題などのタスクではなく，新しいコンテンツを出力する創造的なタスクを処理するモデルを総称して生成AIと呼ぶ．このように，生成AIは，新しい技術を指す表現ではなく，アーキテクチャや能力の大きく異なるモデルを包含する用語となっている．

生成AIの性能は，モデルのサイズ（パラメータ数）の増大とともに向上している．この理由の1つは，一般的な傾向として，モデルのサイズが大きくなるほど，モデルの性能も向上するためである．これをスケーリング則[1]という．2つ目の理由は，モデルの規模が一定水準を超えると，突如として新たな能力が発現するためである．これは創発的能力[2]と呼ばれ，事前に設計されたものではない能力がモデルに自然に現れる現象を指す．たとえば，大規模言語モデルが単なる文章生成を超えて，複雑な推論や問題解決能力を示すようになることが知られている．

しかしながら，生成AIの急速な発展と普及に伴い，さまざまなセキュリティリスクも顕在化してきている．誤回答・誤判断，プライバシー情報の漏洩，反倫理的なコンテンツの出力，なりすましや詐欺への悪用等は，典型的なリスクである．これらのリスクは，従来の情報セキュリティの枠組みでは十分に対処できない新たな課題を提示している．

本稿では，生成AIがもたらすセキュリティリスクと，それに関連する研究動向について包括的に論じる．まず，生成AIの基本原理について解説し，大規模言語モデル（Large Language Model，LLM）と拡散モデルという2つの主要な技術を紹介する．次に，生成AIを安全に利用するための課題を紹介する．さらに，生成AIに対する攻撃手法と，生成AIのサイバー犯罪への悪用可能性について論じる．最後に，これらの課題に対する対策と今後の研究課題について考察する．

## 2. 生成AIの基本原理

生成AIは，基本的には深層学習モデルである．自然言語処理分野ではトランスフォーマーと呼ばれるモデル構造を採用した大規模言語モデル，画像生成分野では拡散モデルが広く用いられている．拡散モデルは，訓練の安定性や複雑なコンテンツの生成に優れ，高品質なデータ生成も可能になっていることから，近年急速に利用が拡大している[3]．敵対的生成ネットワーク（Generative Adversarial Network: GAN）も画像生成に利用されているが，上記の傾向を踏まえて本稿では割愛する．これらのモデルは，多様なタスクに対応できる汎用性の高さから基盤モデル[4]とも呼ばれ，さまざまな応用分野で革新的な成果をもたらしている．

### 2.1 大規模言語モデル

大規模言語モデルは，自然言語を生成する能力を持つ機械学習モデルである．モデルの訓練の中核は，大量の文章データ（コーパス）を用いた，トークン（おおむね単語に相当）の出現確率の推定である．すなわち，トークン列である文の一部を隠して，隠されたトークンを推測させる自己教師あり学習を行う．こうした穴埋め問題を大量に解くことにより，LLMは入力文（プロンプト）に対して自然な文章を生成する能力を獲得する．近年の研究ではLLMが推論能力を有することが判明しているものの，LLMは自然言語を人間のように理解しているわけではなく，あくまで理解したかのような振る舞いを獲得していると考えられる．

モデル構造の点では，遠く離れたトークンの情報を活用して文脈を理解するために，注意機構[5]と呼ばれる仕組みが導入された．この仕組みは，文中の重要な部分に自動的に大きな重みを割り当てるものであり，長文の理解や生成の能力の大幅な向上に寄与した．

### 2.2 拡散モデル

拡散モデル[6]は，画像生成タスクや音声合成タスクにおいて優れた性能を示す機械学習モデルである．その基本原理は，データにノイズを徐々に加えて破壊していく拡散過程と，ノイズを除去して元のデータを復元する逆拡散過程からなる．モデルの訓練時には，ノイズ除去の過程を学習する．すなわち，ノイズにより失われた情報を復元する問題を大量に解く（自己教師あり学習）ことにより，データの再構成能力を高める．生成時にはランダムノイズから出発して逆過程を適用することで，高品質な画像を生成する．この手法により，Stable Diffusion[7]のような画期的な画像生成AIが実現された．

拡散モデルの応用例として，テキストプロンプトからの画像生成がある．ユーザーが文章で指示を与えるだけで，多様で詳細な画像を生成できる．また，既存の画像の編集や，低解像度画像の高解像度化（超解像）にも利用されている．

## 3. 生成AIの安全な利用における課題

生成AIは，さまざまな応用分野で革新的な成果を挙げている反面，セキュリティリスクをもたらす懸念がある．生成AIのセキュリティリスクは，帰納的な計算パラダイムである機械学習モデルに共通するものが多いが，言語や動画像などの複雑なデータを出力する生成AIに特有の論点もある．

まず，機械学習モデルは，人間の意図を越える能力を発揮できる反面，その動作には不確実性を伴うため，外部からの攻撃がなくとも人間と同様に失敗をする．特に，機械学習モデルは生身の身体を通じた感覚や認識を持たないため，記号列である入出力データと実世界の意味とを結びつけることができず，人間では考えられない失敗もする．この問題は記号接地（シンボリック・グラウンディング）と呼ばれる．したがって，機械学習モデルでは，攻撃がない環境でのシステムの安全性を指すセーフティとセキュリティの境界は明瞭ではない．

生成AIの場合には，倫理や法令に違反する文章や動画像，音声を出力するリスクがある．さらに，このリスクに対して，有害な出力を抑制するための対策を施しても，人間のように「説得」に応じて，有害な情報を出力するなど予期せぬ動作をする恐れがある．このような悪意のある「説得」を試みる入力文の探索は，生成AIに対する特有の攻撃と見なせよう．

生成AIは，外部ツールと接続することにより，現実世界に対して直接的かつ相互に影響を及ぼすこともできる．人間に代わって自律的に意思決定するエージェントは典型的な用途である．このような利用方法の場合には，生成AIがもたらす問題はより複雑になる．たとえば，生成AIを搭載したエージェントが，標的となる情報システムの脆弱性検査を行い，自律的に攻撃することも可能かもしれない．

これらの脅威は，セキュリティリスクの概念では必ずしも捉えきれないため，生成AIの特性を考慮した再考が求められる．こうした脅威の悪影響を軽減したものは「信頼できる生成AI」と呼ばれる．

### 3.1 信頼できる生成AIとは

信頼できる生成AIには，安全に利用できること，攻撃への耐性があること，悪用されにくいことが求められる．LLMの信頼性評価について包括的に調査したLiuらの研究[8]によると，信頼できるLLMに必要な要素は，以下の7つの主要カテゴリーに分類される．

- **信頼性**：モデルの出力が一貫して正確で信頼できること．このカテゴリーには，誤情報，ハルシネーション（幻覚）などの問題が含まれる．
- **安全性**：意図しない有害な動作や出力を生成しないこと．機密情報，プライバシー情報，暴力的表現，権利侵害情報など，さまざまな安全上の懸念事項が考慮される．
- **公平性**：特定の集団や個人に対する不当な差別や偏見を含まない出力を生成すること．不公平，ステレオタイプ，選好バイアスなどが関連する．
- **説明可能性と推論能力**：モデルの判断過程や出力の根拠を人間が理解可能なかたちで説明できること．また，論理的な推論能力を持ち，その過程を明示できること．

- **社会規範の遵守**：社会的に受け入れられる規範や価値観に沿った動作をすること．このカテゴリーでは，感情の認識，文化的価値観などが検討される．

- **堅牢性**：攻撃された場合も含めて，入力の変化や予期しない状況に対して堅牢であること．プロンプトインジェクション，データポイズニング攻撃などの問題が含まれる．

- **悪用への耐性**：悪意のある目的で使用されるのを防げること．偽・誤情報やプロパガンダの拡散による社会混乱の誘発や世論操作，マルウェア開発やフィッシング・サイトの構築などのサイバー攻撃が考えられる．そもそも悪意をもって設計または操作されているAIも存在する．

これらの要素を総合的に考慮し，用途や環境に応じて適切なリスク軽減策を講じることが，信頼できる生成AIの実現に不可欠である．次節以降では，これらの要素に関係する生成AIの課題のうち，代表的なものを取り上げて詳述する．

### 3.2　ハルシネーション（幻覚）

ハルシネーション（幻覚）は，生成AIが事実に基づかない情報を生成してしまう現象を指す．ハルシネーションは，原理的に回避が困難であるものの，出力の信頼性と安全性を損ない得るため，生成AIの実用化における大きな障壁の1つとされている．ここでは，LLMのハルシネーションについてのサーベイ論文[9, 10]に基づいて解説する．

#### 3.2.1　ハルシネーションの種類

LLMが対話型AIアシスタントとして進化する中で，ハルシネーションの新たな分類枠組みが提案されている．この枠組みでは，ハルシネーションを事実性ハルシネーションと忠実性ハルシネーションの2つに分類している．

事実性ハルシネーションは，生成されたコンテンツが現実世界の事実と一致しない現象を指す．たとえば，「日本で2番目に高い山は？」という質問に対して，「富士山」と誤情報を生成することが事実の不一致にあたる．また，「ユニコーンの起源」のように，根拠のない事実を捏造するケースも含まれる．

忠実性ハルシネーションは，ユーザーの指示や文脈情報に対して整合性が取れない現象を指し，文脈の不一致，指示の無視，論理的不整合の3つのサブタイプがある．たとえば，ユーザーが「1番好きな食べ物は寿司」と述べた後に，「では，パスタを食べに行きましょう」とLLMが応答する場合が文脈の不一致にあたる．また，指示の無視をする例としては，ユーザーが「50語以内で要約してください」と指示したにもかかわらず，LLMが100語以上の要約を生成する場合が挙げられる．論理的不整合は，推論の過程で矛盾が生じる場合を指す．たとえば，$2x+3=11$を解く過程で，最初のステップでは両辺から3を引いて$2x=8$を得た後，次のステップで両辺を2で割る際に，誤って$x=3$という結果が導き出されてしまう．

#### 3.2.2　ハルシネーションの原因

LLMは，人間のように自然でもっともらしい文を確率的に生成しているに過ぎないため，原理的にハルシネーションは避けられない．また，LLMが知識や事実をどのように記憶しているかも明らかではない点が多い．少なくとも，記号接地は行われておらず，LLMは現実世界の存在を知

らないため，ある事実に関する記述と，これに関するもっともらしくまぎらわしい記述とを区別することはLLMには困難であろう．これ以外のハルシネーションが生じる原因として，以下のようなものが考えられる．

**知識境界**：LLMが獲得できる知識は，訓練データに制約される．この知識量の限界がハルシネーションの原因の1つである．特に，訓練データが最新のものでない場合，最近の出来事や更新された事実に関する質問に答える際にハルシネーションを引き起こす可能性がある．

**訓練データのバイアス**：訓練データに内在するバイアスは，LLMのハルシネーションに大きく影響する．訓練データに重複する情報が多い場合，LLMはその情報に過度に依存するようになり，偏った出力につながる可能性がある．社会的なバイアスが訓練データに内在する場合，LLMはそのバイアスを学習し，ステレオタイプや差別的な表現を出力する可能性がある．

**新しい知識の獲得**：新しい知識を獲得するためのファインチューニングが，既得の知識に関するハルシネーションを引き起こす可能性がある．Gekhmanらの研究[11]では，事前学習で学んでいない知識をファインチューニングで学ぶと，ハルシネーションが増加する傾向が指摘されている．これは，モデルが新しい知識を既存の知識構造と統合する際に矛盾や不正確さが生じるためと考えられる．この現象は，LLMの継続的な学習と更新の過程で特に注意を要する問題であり，効果的な知識統合手法の開発が必要となる．

### 3.3　バイアス（偏見差別）

生成AIにおけるバイアスは，特定のグループに対する偏った，あるいは差別的な表現を出力してしまう現象を指す．Liらの研究[12]では，LLMにおけるバイアスの包括的な調査と分析がなされている．

#### 3.3.1　バイアスの種類

LLMにおけるバイアスとして，便宜的に内在的バイアスと外在的バイアスの2種類の概念が考案されている．

内在的バイアスは，LLMの内部表現，つまり埋め込みに存在するバイアスを指す．これは，大規模なコーパスを用いた事前学習の段階で，言語データに内在する社会的なバイアスがLLMの埋め込みに反映されてしまうことで発生する．たとえば，単語埋め込みにおいて「医師」が男性的概念に，「看護師」が女性的概念に近接するように表現されることがある．このようなバイアスは，モデルのすべての出力に潜在的に影響を与え，ステレオタイプを強化する可能性がある．

外在的バイアスは，LLMが実際に出力する結果に現れるバイアスを指す．これは，内在的バイアスに起因する場合もあれば，特定の下流タスク☆1における訓練データやモデルの設計に起因する場合もある．例として，書類選考のタスクを考える．仮に内在的バイアスが存在しない場合であっても，無意識に男性を優位に評価する採用プロセスから得られた訓練データを利用してファイン・チューニングを行った場合，モデルは男性を有利に評価する外在的バイアスを持つ可能性がある．また，同等のスキルを持つ候補者でも，プログラマーの仕事を男性に，内職の仕事を女性に割り当てる傾向があり得る．これらは，特定の応用分野で直接的な差別を引き起こし，社会的な不平等を助長する可能性がある．

いずれのバイアスも，性別，人種，宗教，年齢，社会経済的背景など，さまざまな属性に関連して現れる．たとえば，特定の人種に対するネガティブな連想や，高齢者の技術的能力に関する先入観なども，LLMの出力に反映される可能性がある．

### 3.3.2　バイアスの要因

LLMにおけるバイアスの主な要因は，訓練データに存在する社会的バイアスである．Liら[12]は，LLMにおけるバイアスの要因を3つの主要なカテゴリに分類している．

**ラベルバイアス**：訓練データのラベル自体に偏りがある場合に発生する．たとえば，ヘイトスピーチや差別的発言を含む未検閲のデータを使用した場合，モデルはこれらの有害な表現を学習し，同様のバイアスを含む出力を生成する可能性がある．また，テキストにラベルを付けるアノテーターの主観が反映され，特定のグループに対する偏見が組み込まれる場合もある．このバイアスは，モデルの基本的な理解や判断基準に影響を与え，広範囲にわたる偏りを生み出す．

**サンプリングバイアス**：訓練データが現実世界のデータ分布を正確に反映していない場合に発生する．特定のジェンダーや人種のデータが過剰または過少に表現されている場合，モデルはその偏りを学習し，不均衡な表現を示す．また，特定の地域や文化に偏ったデータを使用した場合，モデルはグローバルな視点を欠く可能性がある．このバイアスにより，モデルは特定のグループや視点を過度に重視したり，逆に無視したりする傾向を示す．

**意味バイアス**：単語の意味や概念間の関係が社会的なステレオタイプを反映している場合に発生する．特定の職業（例：CEO）と性別（例：男性）の関連付けや，人種と特定の特性（例：知能や犯罪傾向）の関連付けなどがこれに該当する．また，特定の名前や姓が特定の民族や社会的地位と結びつけられる場合もある．このバイアスは，社会的ステレオタイプを強化し，特定のグループに対する偏見を永続化させる可能性がある．

LLMのバイアスを軽減して公平性を確保することは，LLM活用において社会的責任を果たす上で不可欠である．ただし，公平性を含む倫理問題の軽減・解消は次のような難点がある．第1に，訓練データや人間社会に内在するバイアスを軽減・除去するために，こうしたバイアスを知覚することは必ずしも容易ではない．第2に，現実世界を忠実に反映した訓練データを作成しても，倫理的な規範に反するという意味でのバイアスが解消されるわけではない．たとえば，事実として，ほとんど女性が占める職業であっても，女性と推定してはならないとする考え方は，倫理的な規範である．第3に，倫理的な規範は，教育しなければLLMは理解し得ないが，そのためには倫理に反する表現のデータセットも必要になるだろう．こうしたデータセットを安全に作成することも課題である．第4に，倫理的な規範は時代とともに変遷すると考えられるため，公平性の基準は絶えず見直しが必要である．第5に，用途によって倫理的な規範が変わり得る．たとえば，特定の集団の犯罪率が高い統計的事実があった場合に，偏見を助長する用途での表現の利用は許されなくとも，公共政策を立案する場合には有益な情報であり得る．

## 3.4　情報漏洩

生成AIは，非公開の訓練データに含まれるプライバシー情報や機密情報を出力してしまうリスクがある．訓練データに，個人識別情報（PII），企業秘密，個人の信用情報，医療情報などのセンシティブな情報を含める場合には，このリスクに十分な注意を払うべきである．

情報漏洩は，次の2つの場合に分けられる．1つは，訓練データに含まれる情報がモデルに記憶され，直接出力されてしまう場合である．もう1つは，プロンプトインジェクション攻撃などの悪意ある手法により，モデルから故意に情報が引き出される場合である．

情報漏洩は，個人のプライバシー侵害，企業の機密情報流出，これらを通じたサービスへの信頼性低下を引き起こし，法的責任が生じる可能性もある．特に，EUの一般データ保護規則（GDPR）などのプライバシー保護法に違反した場合，厳しい罰則が科される可能性もある．

### 3.5　悪意のあるAIの潜伏

悪意のあるAIの潜伏とは，一見無害に見えるAIシステムが，実際には悪意のある目的のために設計または操作されている状況を指す．このリスクは，データ汚染などを引き起こす攻撃に起因する場合には，次節の生成AIに対する攻撃にも含まれると考えられる．

生成AIの台頭により，AIシステムが処理する入出力情報は質的量的に変化し，適用範囲が拡大した．これと同時に，膨大な数の生成AIが公開されるようになった．この反面，悪意のあるAIを検出するコストも上昇しており，悪意のあるAIの潜伏問題はますます深刻化している．これに伴いAIシステムに求められる信頼性を確保するための要求水準も上がっている．

悪意のあるAIは，次のような形態で開発・公開されると考えられる．1つ目は，訓練データの汚染によるモデル機能の改変である．たとえば，顔認識システムに特定の人物を誤認させるバックドア☆2や，マルウェア検出システムに特定のマルウェアを見逃すようなバックドアを仕込むことも考えられる．2つ目は，生成AIを使ったサービスが，不正な情報収集に悪用される場合である．ディープフェイクを作成するアプリケーションの入力情報を使ってフェイク・コンテンツが作成され，詐欺☆3や脅迫に悪用される可能性がある．3つ目は，プライベートな相談などを行うチャットツールにおいて，犯罪や自殺を教唆するなどの場合である．これらによって精神に有害な影響を及ぼす可能性が考えられる．

正常なAIと悪意のあるAIの区別は容易ではないが，モデルの安全性を客観的に検査・評価する枠組みを整備することが望ましい．特に，生命や財産にかかわる医療や金融などの高リスク分野での活用や，未成年者の利用を念頭に置くと，安全性を検証する重要性は高いと考えられる．

## 4．生成AIに対する攻撃

生成AIに対する攻撃は，著作権侵害や個人情報漏洩，有害情報の生成などのリスクをもたらす．Shayeganiらの研究[13]では，LLMに対する攻撃手法と防御手法について論じている．前章では，主として攻撃がないもとでの生成AIのリスクを論じたのに対して，本章では攻撃に起因するリスクを論じる．

### 4.1　Jailbreak

Jailbreak攻撃は，生成AI，特にLLMの安全対策を回避し，本来は生成すべきではない出力を強制的に生成させる攻撃手法である．Weiら[14]とAnilらの研究[15]では，Jailbreak攻撃の仕組みや，その危険性について解説している．Jailbreak攻撃は，大きく分けて2つの段階からなる．

**安全対策の回避**：LLMは，倫理的に問題のある出力や，有害な情報を生成しないように，さまざまな安全対策が施されている．Jailbreak攻撃の第1段階では，特殊なプロンプトを入力することで，これらの安全対策を回避する．たとえば，Weiらの研究[14]では，**Competing Objectives**と呼ばれる手法が紹介されている．これは，LLMが本来持つ言語モデルとしての能力と，安全対策の要件との間に矛盾を生じさせることで，安全対策を無効化する手法である．

**不正な出力の生成**：安全対策を回避した後，攻撃者は，LLMに対して，本来は生成すべきではない出力を強制的に生成させる．たとえば，差別的な発言，個人情報，マルウェアのソースコードなどを生成させることが考えられる．

Anilら[15]は，多数の有害な質問と回答のペアをLLMに文脈内学習☆4させることで，Jailbreak攻撃が成功することを報告している．また，LLMのコンテキストウィンドウ☆5が大きくなるにつれて，その効果が増すことが示唆されている．この手法は，**Many-shot Jailbreaking**と呼ばれる．

### 4.2 データポイズニング攻撃

データポイズニング攻撃は，機械学習モデルの訓練データに悪意のあるデータを意図的に混入させ，モデルの動作を改変する攻撃手法である．この攻撃の目的は多様で自由度が高い．目的として，モデルの全体的な精度低下，有害情報や機密情報の出力，攻撃者があらかじめ指定した特徴を持つ入力に対してのみ挙動を変化させるバックドアの設置などが考えられる．たとえば，顔認識システムの訓練データに，特定の人物を誤って分類するように加工された画像を混入することで，その人物の識別のみを意図的に失敗させることが可能となる．データポイズニング攻撃は，主に以下の方法で行われる．

- **ラベル改ざん**：訓練データの正解ラベルを不正なものに変更し，モデルの精度を低下させる．
- **データ注入**：悪意のあるデータを訓練データセットに追加し，モデルの動作を改変する．
- **特定パターンの埋め込み**：特定の入力パターンに対して，意図した出力を生成するようモデルを誘導する．

特に，生成AIは複雑な入出力データを取り扱うことができるため，通常の機械学習モデルよりも複雑な論点を抱え得る点には留意が必要である．たとえば，Shanら[16]が提案するGlazeは，アーティストがわずかなノイズを自身の作品に加えてインターネットに公開するものである．アーティストの許諾を得ずに，ノイズ入りの作品を訓練データに加えてモデルの訓練を行うと，ノイズの付加が一種のデータポイズニング攻撃として働く．訓練済みモデルを利用して，当該作品に類似した画像を生成しようとすると，生成AIは混乱して無意味な画像を生成してしまう．提案者らは，この技術が著作物の無断利用からのアーティストの権利保護に役立つとしているが，こうしたデータの公開には賛否両論がある．

## 5. 生成AIのサイバー犯罪への悪用

生成AIの多言語展開能力，マルチ・モーダルな展開能力は，フィッシング攻撃，マルウェア等の開発，フェイク・コンテンツを悪用した詐欺と世論操作などのサイバー攻撃に悪用され得る．

### 5.1 フィッシング攻撃

生成AIは，文法の修正や翻訳の精度が高く，自然な文章の例示もできることから，フィッシングメールの作成に悪用できるとの指摘がある[17]．標的を特定の人物に絞らない標準的なフィッシング攻撃については，すでに攻撃手順が相当に自動化されている．このため，生成AIの出現による追加的な脅威の高まりは限定的であると評価されている．この一方，標的を特定の人物や組織に絞るスピアフィッシング攻撃では，生成AIを利用して，より洗練された攻撃が可能になる恐れがある．スピアフィッシング攻撃は，特定の人物や組織を標的とし，なりすまし等によって騙すことにより，企業秘密やパスワードなどの秘密の情報を聞き出すことを目的とする．標的に固有の情報を調査する必要があるため，この攻撃は労働コストが高く規模拡大が困難である．標的の信用を得るために，エージェントベースのLLMを使い，双方向の会話を行えれば，攻撃を洗練させて成功確率を引き上げることができるかもしれない．また，声の模倣や架空の人物のプロフィールを精巧に偽造することで，特定の人物または架空の人物へのなりすましが自動化され，コスト低下と大規模化を可能にする恐れがある．既存研究[18]では，標的となる個人またはグループに関する固有の情報を交えたフィッシングメールは，生成AIで作成したものでも情報窃取の成功率が高いことが示されている．

また，生成AIは，攻撃基盤となるフィッシング・サイトの構築にも有用であることが示されている[19]．商用のLLMを使った実験では，フィッシング・サイトの構築を直裁に命じるプロンプトは受理されないものの，特定のブランドに模したWeb・サイトの作成，データ入力フォームの設置，入力情報の攻撃者への送信などの機能単位に分けるとプロンプトが受理されることが判明している．生成AIは，攻撃基盤の構築に要するコストやスキルの面で参入障壁を下げると見られる．

### 5.2 マルウェア等の開発

生成AIは，情報窃取マルウェアやランサムウェアなどのマルウェア開発，ダークウェブの取引サイトの開発にも悪用できると指摘されている[17]．LLMの強力なコーディング支援機能は，マルウェアの開発にも悪用できるため，同開発に要求されるスキルのレベルを引き下げると考えられる．また，既存研究[20]では，感染端末でプログラムが実行されるたびに，LLMにAPI経由でアクセスすることで，キーロガー☆6を作成・実行できることが示されている．その際，プロンプト実行ごとに，コードの微妙に異なるバージョンが生み出されるため，マルウェア検知を回避しやすいとのリスクが指摘された．

### 5.3 フェイク・コンテンツを悪用した詐欺と世論操作

生成AIを用いて，架空または特定の人物に関する精巧な偽物の画像・動画・音声を作成できる．この技術を善良な目的に利用すれば，映画制作において，代役の顔に本物の俳優の顔を埋め込んだ動画を合成するなどの商業利用が可能である．この反面，架空または特定の人物になりすまして詐欺に悪用すること，オンラインでの本人確認であるeKYC（electronic Know Your Customer）の突破，などに悪用することもできてしまうと考えられる．

SNSを通じた偽・誤情報の拡散による社会混乱の誘発，世論誘導にもフェイク・コンテンツは悪用されていると言われる．先進各国において選挙活動のオンライン化が進み，外国勢力の介入も容易になっていると見られる中，こうした脅威は増大していると考えられる[21]．

## 6. 技術的なリスク軽減策

これまで生成AIの課題とリスクについて述べてきたが，本章では生成AIがもたらすセキュリティリスクの軽減策のうち技術面を解説する．大別すれば，生成AIそのものの安全性を高めるアプローチと，生成AIの外部装置で安全性を担保するアプローチがある．

### 6.1 アラインメントの強化

アラインメントとは，モデルが人間の嗜好に合わせた出力をするように調整することを指す．アラインメントの強化は，主にハルシネーションとバイアスのリスクを軽減する．たとえば，OpenAIはGPT-4のリリース前に6カ月をかけてアラインメントの反復改善を行っている[22]．

具体的なアラインメント手法としては，強化学習による人間のフィードバック（RLHF: Reinforcement Learning from Human Feedback）[23]や直接選好最適化（DPO: Direct Preference Optimization）[24]が挙げられる．RLHFは人間の評価者からのフィードバックを基にモデルの出力を調整する手法であり，DPOは人間の選好データを直接モデルに学習させる手法で，RLHFよりも計算効率が高いとされる．

### 6.2 ガードレールの実装

ガードレールとは，生成AIが安全に，かつ意図したとおりに動作するようにするための一連の対策のことである[25]．ガードレールの主な目的は，有害な内容や不適切な情報の生成を防ぎ，個人情報や機密情報の漏洩を防止し，AIシステムの悪用や攻撃を防ぐとともに，AIの判断や行動が人間の価値観や倫理観に沿うようにすることである．

これらの目的を達成するため，ガードレールはAIシステムにおいて，さまざまなコンセプトやアプローチに基づいたものが実装される．これらの対策は，以下の3つのカテゴリ（層）に便宜的に分類できる．

- **ゲートキーパー層**：生成AIの外部に位置し，入力されるプロンプトや指示を検査する層である．この層では，悪意のあるプロンプトやプロンプトインジェクション攻撃を検出し，ブロックする．システムの制御を奪おうとする不正なコマンドや，個人情報を抽出しようとする試みを防ぐ役割を果たす．
- **ナレッジアンカー層**：生成AIの知識基盤を強化し，出力の信頼性を高める層である．この層では，外部の信頼できる知識ソースとLLMを連携させ，最新かつ正確な情報に基づいた応答を生成する．たとえば，Retrieval-Augmented Generation（RAG）技術を用いて，生成AIの回答を外部データベースの情報で補強する．
- **パラメトリック層**：生成AIの内部パラメータや学習プロセスを直接制御する層である．この層では，モデルの挙動をより安全で信頼性の高いものに調整する．バイアスを軽減するためのファインチューニングや，プライバシーを保護するための差分プライバシー技術の

適用などが行われる.

これらの多層的な対策を組み合わせることで，より安全で信頼性の高いAIシステムを構築することができる．たとえば，医療診断支援AIでは，ゲートキーパー層で患者の個人情報を保護し，ナレッジアンカー層で最新の医学知識を参照し，パラメトリック層で公平性を確保するなど，複数のセーフティガードが連携して機能する.

### 6.3　データセットの改善

データセットの改善は，多くのリスクを軽減するための重要な対策である．高品質で多様性のあるデータセットを用いることで，AIモデルの性能向上と同時に，さまざまなリスクの軽減が可能となる.

データセット改善の具体的なアプローチには以下がある.

- **データの多様性確保**：さまざまな属性や背景を持つデータを収集し，バイアスを軽減する.
- **データクリーニング**：不正確または不適切なデータを除去し，ハルシネーションのリスクを低減する.
- **プライバシー保護データの利用**：個人情報を適切に匿名化または仮名化したデータを使用し，情報漏洩リスクを軽減する.
- **定期的なデータ更新**：最新の情報を反映させることで，時間経過によるハルシネーションを防ぐ.

これらの取り組みにより，AIモデルの学習基盤となるデータセットの質が向上し，多くのセキュリティリスクの軽減が期待される.

## 7. まとめと今後の課題

本稿では，生成AIがもたらすセキュリティリスクと研究動向について概観した．生成AIは，その高度な能力からさまざまな産業分野で恩恵をもたらす一方で，ハルシネーション，機密情報の漏洩，有害情報の出力，偽・誤情報の作成やサイバー犯罪への悪用など，多岐にわたる脅威ももたらす.

これらの脅威を完全に取り除くことは原理的に困難であるが，リスクを軽減することは可能である．技術的なリスク軽減策には，生成AIそのものの安全性を高めるアプローチと，生成AIの外部装置によって安全性を確保するアプローチがある.

リスク軽減を図っていく上で，今後の課題として以下の論点が指摘できる．第1に，生成AIの安全性を検査し，評価する枠組みの確立が求められる．オープンソースで膨大な数の生成AIが開発されて利用できる状態にあるが，これらのモデルを安全に活用するためには，プログラム・ソースコードと同様に信頼性の評価が不可欠である．第2に，法規制を中心とする社会制度の整備も生成AIの活用深化に不可欠である．著作権や知的財産権，プライバシー権，責任の所在の明確化などの法的課題のほか，公平性，反差別，透明性と説明可能性の確保などの倫理ガイドライ

ンの策定も重要となる．また，偽・誤情報の流通への対策として，コンテンツの発信者を証明するオリジネータープロファイルなどの制度整備も有用な選択肢となり得る．第3に，生成AIの不確実性や悪用への対策として，一般利用者のリテラシー教育も重要である．生成AIのもたらす脅威に技術的・制度的な対策を施しても，すべてのリスクを取り去ることはできないからである．

生成AIの恩恵を享受するためには，生成AIのもたらすリスクを軽減しつつ，残余リスクを受容することが肝要である．また，倫理的な規範は国や文化圏に応じて異なるため，日本の国情に合った生成AIを利用するために，日本が独自に生成AIを開発する努力も有益であろう．生成AIの社会での活用を深めていくためには，研究者，開発者，政策立案者，そして一般利用者の理解深耕と協力が不可欠であり，本稿がそのための一助となれば幸いである．

## 参考文献


1）Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D.: Scaling Laws for Neural Language Models, *arXiv preprint arXiv: 2001.08361* (2020).
2）Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent Abilities of Large Language Models, *Transactions on Machine Learning Research* (2022).
3）岡野原大輔：拡散モデルデータ生成技術の数理，岩波書店（2023）．
4）Bommasani, R., Hudson, D. A., Adeli, E., Altman, R. B., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the Opportunities and Risks of Foundation Models, *arXiv preprint arXiv:2108.07258* (2021).
5）Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* (2015).
6）Ho, J., Jain, A. and Abbeel, P.: Denoising Diffusion Probabilistic Models, *Advances in Neural Information Processing Systems*, Vol.33, pp.6840–6851 (2020).
7）Stability AI Japan（株）：画像生成AI Stable Diffusion，https://ja.stability.ai/stable-diffusion (Accessed 31-08-2024).
8）Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F. and Li, H.: Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment, *arXiv preprint arXiv: 2308.05374* (2023).
9）Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A. and Das, A.: A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, *arXiv preprint arXiv: 2401.01313* (2024).
10）Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, *arXiv preprint arXiv: 2311.05232* (2023).
11）Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R. and Herzig, J.: Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?, *arXiv preprint arXiv: 2405.05904* (2024).
12）Li, Y., Du, M., Song, R., Wang, X. and Wang, Y.: A Survey on Fairness in Large Language Models, *arXiv preprint arXiv: 2308.10149* (2023).
13）Shayegani, E., Mamun, M. A. A., Fu, Y., Zaree, P., Dong, Y. and Abu-Ghazaleh, N.: Survey of Vulnerabilities in Large Language Models Revealed by Adversarial

Attack, *arXiv preprint arXiv: 2310.10844* (2023).
14）Wei, A., Haghtalab, N. and Steinhardt, J.: Jailbroken:How Does LLM Safety Training Fail?, *arXiv preprint arXiv: 2307.02483* (2023).
15）Anil, C., Durmus, E., Sharma, M., Benton, J.,Kundu, S., Batson, J., Rimsky, N., Tong, M., Mu, J., Ford, D., et al.: Many-shot Jailbreaking, *Proceedings of the 38th Annual Conference on Neural Information Processing Systems* (2024).
16）Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R. and Zhao, B. Y.: Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models, *arXiv preprint arXiv: 2302.04222* (2023).
17）Bezzi, M.: Large Language Models and Security, *IEEE Security & Privacy*, Vol.22, No.2, pp.60–68.
18）Bethany, M., Galiopoulos, A., Bethany, E., Karkevandi, M. B., Vishwamitra, N. and Najafirad, P.: Large Language Model Lateral Spear Phishing: A Comparative Study in Large-Scale Organizational Settings, *arXiv preprint arXiv: 2401.09727* (2024).
19）Roy, S. S., Thota, P., Naragam, K. and Nilizadeh, S.: From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models, *Proceedings of the 2024 IEEE Symposium on Security and Privacy*, pp.36–54 (2024).
20）Sims, J.: BlackMamba: Using AI to Generate Polymorphic Malware, https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware (Accessed 19-08-2024).
21）Cybersecurity and Infrastructure Security Agency (CISA): Risk in Focus: Generative A.I. and the 2024 Election Cycle.
22）Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 Technical Report, *arXiv preprint arXiv: 2303.08774* (2023).
23）Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training Language Models to Follow Instructions with Human Feedback, *Advances in Neural Information Processing Systems*, Vol.35, pp.27730–27744 (2022).
24）Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S. and Finn, C.: Direct Preference Optimization: Your Language Model Is Secretly a Reward Model, *Advances in Neural Information Processing Systems*, Vol.36, pp.53728–53741 (2023).
25）Ayyamperumal, S. G. and Ge, L.: Current State of LLM Risks and AI Guardrails, *arXiv preprint arXiv: 2406.12934* (2024).

**脚注**

☆1　汎用性の高い基盤モデルの訓練を上流タスクと捉え，これと対比して，基盤モデルを所与に特定のタスクの処理に向けて行うファインチューニングを下流タスクと呼ぶ．
☆2　攻撃者があらかじめ指定した特徴（トリガー）を入力データが持つ場合に，機械学習モデルの挙動が変化するような隠蔽された機能をバックドアと呼ぶ．
☆3　2024年2月に英Arup社が，ディープフェイクを悪用した詐欺により，2億香港ドル（約37億円）を騙し取られる事件が発生した．この事件では，同社の財務部門の最高責任者の音声付き動画のディープフェイクが作成され，ビデオ通話において，財務部門の従業員が不正な資金移動を命じられた．https://www.cnn.com/2024/05/16/tech/arup-de

epfake-scam-loss-hong-kong-intl-hnk

☆4　in-context learning：関連する情報をプロンプトに含めることで，事前知識をモデルに与える手法．

☆5　LLMに入力可能な最大トークン数．

☆6　キーボード入力を記録するマルウェアの一種．

**染谷実奈美**
someya@ai.iisec.ac.jp
2016年名古屋大学理学部物理学科卒業．2018年より官公庁にて情報セキュリティ業務に従事．2024年情報セキュリティ大学院大学情報セキュリティ研究科博士前期課程修了．同大学同研究科博士後期課程在学中．

**菅　和聖**（正会員）
kazutoshi.kan@boj.or.jp
日本銀行金融研究所情報技術研究センター企画役．2007年東京大学工学部計数工学科卒業．2009年東京大学大学院情報理工学系研究科数理情報学専攻修了．2009年日本銀行入行．本会CSEC研究会会員．人工知能学会 安全性とセキュリティ研究会主幹事．

**大塚　玲**（正会員）
otsuka@iisec.ac.jp
情報セキュリティ大学院大学教授．2002年東京大学大学院工学系研究科電子情報工学専攻博士課程修了．博士（工学）．2005年より2017年まで産業技術総合研究所．2017年より現職．本会シニア会員．電子情報通信学会BioX研究会顧問．人工知能学会 安全性とセキュリティ研究会主査．

投稿受付：2024年9月9日
採録決定：2024年11月18日
編集担当：南山泰之（国立情報学研究所）