4 生成 AI のセキュリティ

生成AIのセキュリティリスクとその対策、サイバーセキュリティ分野への影響

個人、企業を問わず生成 AI の利用が広がっている。便利な一方で無視できない生成 AI のセキュリティリスクと、注意すべきポイントを紹介する。さらに、生成 AI がサイバーセキュリティの世界をどのように変えていく可能性があるか考察する。

株式会社 NTT データグループ
技術革新統括本部 システム技術本部
サイバーセキュリティ技術部
課長代理　**堰根 哲平** 氏

生成 AI とセキュリティ

ChatGPT の登場以来、現代の産業革命とも言われる生成 AI は、世界中の関心を惹きつけている。しかし、新たな技術には新たなセキュリティリスクが付き物である。生成 AI においても、技術が持つリスクを正しく認識して利用することが重要である。さらに、生成 AI は高度な汎用性を持つため、生成 AI 自体のセキュリティリスクにとどまらず、サイバーセキュリティ分野全体に影響を及ぼす可能性も無視できない。

なお、生成 AI という名称は様々な意味で使用されることがあるが、本稿では ChatGPT を代表とする大規模言語モデル（以下 LLM）を利用した対話型 AI に対象を限定する。それ以外の生成 AI（画像生成 AI など）に言及する場合は明記する。なお、ここでは生成 AI に特有のリスクのみを扱うこととし、例えば DoS 攻撃やサプライチェーン攻撃のような、生成 AI の利用有無にかかわらず一般的な情報システムで想定されるリスクは本稿では割愛する。

ユーザーが注意すべきセキュリティリスク

最初に、ChatGPT のようなサービスを利用するユーザーが注意すべきセキュリティリスクについて考える。ユーザーと生成 AI の間では、ユーザーが「プロンプト」と呼ばれる命令を入力し、生成 AI は受け取った命令にふさわしい（と AI が算出した）回答を出力する。

ここで想定される主なリスクは入力情報の漏洩と、不適切な情報の出力の 2 つである。

まず入力情報の漏洩であるが、ユーザーが生成 AI に入力する情報は、生成 AI の学習に使用される場合があり、入力情報がほかのユーザーへの出力に含まれる形で漏洩する可能性がある。対策は、漏洩を避けるべき情報を入力しない、または入力情報を学習に使用しない、かつ入力情報が適切に扱われることが保証されたサービスを利用することである。

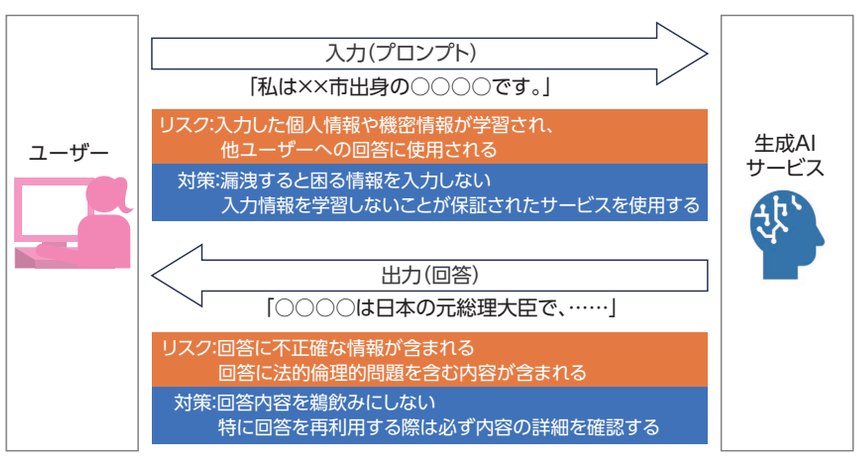

図 1　生成 AI サービスのユーザーから見たセキュリティリスク

2つ目のリスクは、生成AIの不適切な情報の出力である。生成AIは事実と異なる内容をもっともらしく出力する、つまり平然と嘘をつく可能性がある。また、著作権侵害や差別、偏見など法的、倫理的な問題を含む出力をする可能性もある。これらの不適切な出力を減らすため、生成AIの開発者が対策に取り組んでいる。しかし、現状完全に無くすことは困難であり、ユーザーが生成AIの出力にこれらのリスクがあることを認識し、鵜呑みにしないことが大切である。特に、生成AIの出力を参考にレポートを作成するなど出力を再利用する場合は、その正確性やコンプライアンスはユーザーが責任を持ってチェックしなければならない。

生成AIをサービスに組み込む場合のリスク

次に、生成AIを組み込んだサービスの提供者から見たセキュリティリスクを考える。以前からユーザーの一時対応窓口としてチャットインターフェイスを設けるケースがあったが、ここに生成AIを利用するのが典型的なユースケースである。ここで想定されるリスクは、ユーザーからの予期しない入力に対して、不適切な出力を行うことである。入力と出力について順に詳細を見ていく。

入力としては、一般的なユーザーだけでなく、悪意を持ったサイバー攻撃者からに狙われる可能性も想定される。サイバー攻撃者は、通常では発生し得ない不適切な出力を意図的に発生させるため、細工したプロンプトを入力する。このようなプロンプトは「敵対的プロンプト」と呼ばれている。なお、ユーザーからの入力を生成AIが学習するサービスでは、政治的や倫理的に偏った入力を大量に受けた後に、出力に不適切なバイアスがかかるリスクについても考慮が必要である。

次に出力であるが、敵対的プロンプトを含む入力を受け付けた場合の出力には、大別して3つの問題が発生するリスクがある。

- ●他ユーザーの情報やシステムの内部情報などを出力する情報漏洩
- ●事実誤認、偏見や差別、犯罪の教唆、著作権侵害などのサービスとして不適切な回答
- ●生成AIの出力が連携システムの入力として使用される場合に、連携システムの意図しない動作の惹起

これらの共通的な対策として、前節で触れた生成AI開発者が用意している不適切な回答を抑制する仕組み（ガードレール）に加えて、以下に列挙する対策を多層的に行うことを検討する必要がある。

- ●サービス内で追加設定するプロンプトによる挙動の制限（特定の領域以外の質問には回答しないなど）
- ●入出力の形式的なチェック（出力は“Yes”か“No”のみ許可するなど）
- ●ユーザーとの入出力を行う生成AIとは異なる独立した生成AIによる入出力のチェック
- ●敵対的プロンプトの実施を含む脆弱性試験
- ●運用フェーズにおいて生成AIや連携システムに不審な挙動がないか監視

なお、対策を強固にするほど入出力の自由度が下がるため、サービスの性質によって何をどこまで制限するのかの見極めが求められる。リスク許容度が低いサービスでは、本当に生成AIを利用すべきか十分な検討が必要となる。

国際的なアプリケーションセキュリティの団体であるOWASPがLLMを利用したアプリケーションの特に重大なセキュリティリスクを解説した“OWASP Top 10 for LLM Applications”を発行している。本稿で割愛した生成AIに特有でないリスクも取り上げられており、参考になるだろう。OWASP Japanによる日本語訳も公開されている。

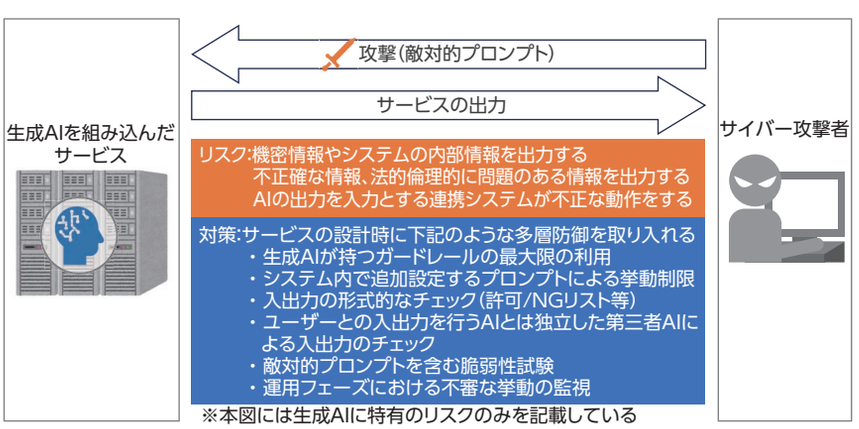

図2　生成AI利用サービスの提供者から見たセキュリティリスク

生成 AI がサイバーセキュリティ全体に与える影響

ここまで生成 AI 自体のセキュリティリスクと対策について論じてきた。本節ではそこから離れて、生成 AI がサイバーセキュリティ全体に与える影響について考える。これまでの一般的なソフトウェアに比べて、生成 AI は利用のハードルが低い一方で、極めて汎用性が高い特徴がある。そのため、生成 AI の発展はサイバーセキュリティにも様々な形で影響が及ぶと考えられる。

本稿執筆時点において知りうる限り、生成 AI を利用した全く新しいサイバー攻撃が出現したという報告はないが、サイバー攻撃者の一部がすでに生成 AI を利用していることは間違いない。下記のように既存のサイバー攻撃の高度化と効率化に悪用されていると推測される。

- **詐欺への悪用**：フィッシングや BEC（ビジネスメール詐欺）を目的としたメールやフェイクニュースの生成（この用途では本稿で対象としている対話型 AI だけでなく、画像や音声、動画の生成 AI も効果的に使用される）
- **マルウェア作成支援**：現状の生成 AI だけで高度なマルウェアを作ることは難しいが、スキルを持つ攻撃者が作成効率を上げるためのプログラミング支援ツールとして悪用
- **調査・偵察支援**：攻撃前の調査・偵察行為において、人間が行っていた判断の自動化や効率化に悪用

これらの反社会的な用途は、本来ならば生成 AI に組み込まれたガードレールにより回答が拒否されるべきであるが、攻撃者はガードレールの回避を狙った敵対的プロンプトを洗練させてきた。生成 AI の開発元が制限を厳しくすると、今度は一切の出力制限がなくサイバー攻撃向けのチューニングが施された攻撃用生成 AI が登場し、闇ビジネスとしてサイバー攻撃者に販売されている。

一方で、生成 AI の恩恵を受けるのはサイバー攻撃者だけではない。CSIRT や SOC もサイバー攻撃に対抗する手段として、生成 AI の積極的な活用が求められるようになるだろう。未来には、完全自動化されたサイバー攻撃を、完全自動化された防御システムが止めるような光景も見られるかもしれない。しかし、今の時点で生成 AI にすべての対応を完全に任せることは難しく、当面は下記のような生産性を向上し、人的リソースの不足を補うような用途が中心になると考えられる。

- **ツール作成支援**：CSIRT や SOC の日常的な業務や、インシデント調査・対応作業を効率化するためのツール作成の支援
- **脆弱性調査**：攻撃者の調査・偵察支援の裏返しとして、自組織の攻撃を受ける可能性がある箇所（アタックサーフェス）の調査を効率化
- **アシスタント**：インシデント調査・対応作業のアシスタントとして作業を支援。的確な指示（プロンプト）を与えるため、利用者に一定のスキルが必要

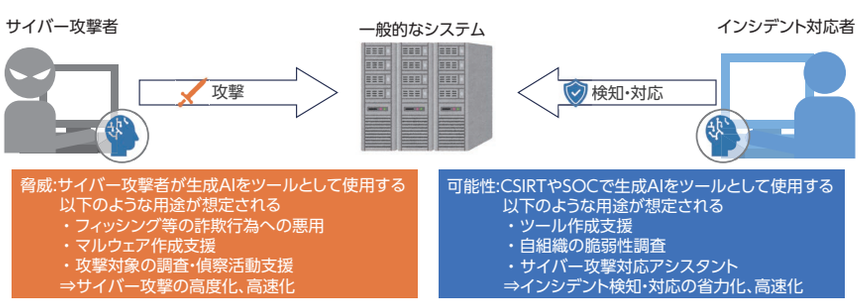

図 3　生成 AI がサイバーセキュリティにもたらす脅威と可能性

NTT DATA の取り組み

NTT DATA では、グローバルのセキュリティ関連組織が連携して生成 AI のセキュリティについて調査や検証を進めている。

特に力を入れているのが、生成 AI をセキュリティ運用に活用する取り組みである。高度セキュリティ人材の不足が世界規模の課題となる中で、セキュリティ運用における生成 AI の活用は必須になると考えられる。取り組みの具体例として、脅威ハンティングにおける調査クエリの自動生成や、セキュリティインシデント検知における誤検知削減、SOC アナリスト向けのアシスタントなどの検証を行っている。検証では、生成 AI に過去のインシデント対応情報を参照させて、精度を向上させる試みも実施している。

これらの取り組みの成果を積み重ねていくことで、NTT DATA が持つグローバルのセキュリティナレッジを集約した生成 AI を用い、インシデントへのプロアクティブな対応の実現を目指している。