調査・分析レポート

情報操作型サイバー攻撃の脅威（３）
―生成 AI による情報操作―

独立行政法人情報処理推進機構　サイバー情勢研究室　研究員　**長迫　智子**

I. はじめに

近年、世界各国で国家支援型サイバー攻撃が激増しており、日本においても港湾や病院などの重要インフラや学術機関などが攻撃を受けている。さらには、機能破壊型や情報窃取型のサイバー攻撃にとどまらず、世論操作、社会分断を企図する情報操作型のサイバー攻撃も増加しており、既存類型のサイバー攻撃と融合して実行されるハイブリッド型のサイバー攻撃へと高度化、複雑化が進んでいる。2016年の米国大統領選を契機として、各国の選挙における情報操作型サイバー攻撃による影響工作がみられ、またウクライナ戦争やイスラエル・ハマスの武力衝突といった有事においても、サイバー空間を通じた情報戦、認知戦が繰り広げられた。

2024 年５月号からの連載[1] にて、情報操作型サイバー攻撃をめぐる情勢とそれを用いた情報戦の様相を整理、議論してきたが、本連載の３回目である本稿では、生成 AI を用いた攻撃事例が増えていることに焦点を当てる。新たな技術である生成 AI がわれわれの社会のトラストをどのように毀損しているか、技術的な概略を述べたうえで、生成 AI を悪用した情報操作型サイバー攻撃の事例を分析し、そのような攻撃に対してどのように対処すべきか、その対策を議論する。

II. 生成 AI の脅威

１．生成 AI とは何か

AI とは「人工知能（Artificial Intelligence）」の略であるが、それがどのようなモデルを利用しているかによって更に分類されることとなる。そのうちの一つが「生成 AI（Generative AI）」である。生成 AI とは、テキストや画像、映像などの生データ（例えば、シェイクスピアの作品テキストやレンブラントの絵画集など）を取り込み、指示に基づいた出力を促されたときに統計的に可能性の高い出力を生成するように「学習」できる「ディープラーニング（Deep Learning：深層学習）」モデルを利用しているものを指す。より高度な生成 AI では、生成モデルは学習データの単純化された表現をパターン化し、そこから学習元のデータに似ているがそれとは同一ではな

[1] 長迫智子「情報操作型サイバー攻撃の脅威（1）　―ディスインフォメーションを利用した情報戦の現状と課題―」『CISTEC Journal』No.211（2024 年 5 月）、155 頁―163 頁。
長迫智子「情報操作型サイバー攻撃の脅威（２）　―第 6 の戦場としての認知領域―」『CISTEC Journal』No.212（2024 年 5 月）、167 頁―177 頁。

い全く新しい作品を作成することができる[2]。生成モデルは、統計学において数値データを分析するために長年使用されてきた。しかし、ディープラーニングの進化により、画像や音声、その他の複雑なデータタイプに拡張することが可能になった。

このような生成 AI を利用したウェブサービスも増加しており、代表的なサービスは以下の通りである。

・テキスト生成 AI（プログラミングのコード生成にも応用可能）
ChatGPT、Claude 3、Gemini など
・画像生成 AI
Stable Diffusion、Midjourney など
・音声生成 AI
Text-to-Speech AI、VALL-E など
・音楽生成 AI
Suno AI、MusicLM など
・動画生成 AI
Sora、Runway Gen-2 など

これらは無料で利用を開始できるものも多く、より高度なメニューを利用するためには有料会員登録等が必要となるが、特別なスキルやソフトを有しない一般人でも気軽に生成 AI を利用することができる環境となっている。生成 AI が身近なツールとなったことによる利点は枚挙に暇がないものの、健全な情報空間や社会のトラスト形成という側面からは有害な面もあり、現時点では解消できていない問題も多い。

情報空間の偽・誤情報に関連する一例としては、「ハルシネーション（Hallucination：幻覚）」が挙げられる。ハルシネーションとは、もっともらしいが「事実とは異なる内容」や「文脈と無関係な内容」といった誤情報を AI が生成することである[3]。人間が現実の知覚ではなく脳内の想像で「幻覚」を見る現象と同様に、まるで AI が「幻覚」を見て出力しているように見えるため、このように呼ばれる。現在の生成 AI は、実際にはトレーニングしていない情報をあたかも「幻聴／幻視」しているように、信頼できない出力や誤解を招く出力を生成する場合がある。しかし、出力を受け取ったユーザ側にその正誤を判断できる知識がない場合、さらには AI を SF めいた万能のシステムだとそのユーザが誤認しているような場合には、「AI が出した正しい答え」として SNS 等で拡散されてしまう例が散見される。そのため、生成 AI サービスを利用するには、ユーザ側にも一定程度のリテラシーが求められる。その他にも、学習データの著作権や倫理上の問題等、生成 AI にはいくつかの問題があるが、本連載の主題は情報戦下の情報操作型サイバー攻撃であるため、こうした攻撃に関連する「コンテンツの大量生成」と「ディープフェイク」の二点に焦点を当てる。

２．生成 AI の脅威

（１）コンテンツの大量生成の悪用

生成 AI は、学習用に入力されたデータを一定のパターン化、記号化により学習し、新たなコンテンツを生み出すことが出来る。例えば、大量のレンブラントの絵画をデータとして学習させ、レンブラントの作風を踏襲した新作を現代に創出することができる[4]。また、こうした技術を応用して、過去の作品の欠損部分を復元するといったことも可能となった[5]。

単なるコピーアンドペーストではなく、一定の傾向、パターンのもとでリアリティを有した差分あるコンテンツを大量に生成するという技術は、SNS 上での影響工作にも悪用されている。

その一つが、AI によるボット（ソーシャルメディアボット。SNS 上でロボットのように自動で投稿を行うアカウントのことを指す。）の大量生成である。これまでの各国の選挙等における影響工作では、戦略的ナラティブやディスインフォメーションを拡散するためにボットアカウントが活用されてきた。し

[2] Kim Martineau, “What is generative AI?,” IBM Research, 20 Apr 2023. (https://research.ibm.com/blog/what-is-generative-AI)
[3] 一色政彦「ハルシネーション（Hallucination）とは？」『AI・機械学習の用語辞典』IT media、2024 年 3 月 4 日。（https://atmarkit.itmedia.co.jp/ait/articles/2303/30/news027.html）
[4] EMILY REYNOLDS, “This fake Rembrandt was created by an algorithm,” WIRED, 7 APR 2016. (https://www.wired.com/story/new-rembrandt-painting-computer-3d-printed/)
[5] 産経新聞「レンブラントの傑作「夜警」、ＡＩで復活」2021 年 6 月 24 日。（https://www.sankei.com/article/20210624-2K7BOFTYCRMIBLW7TRE5ZJR6RU/）

かし、従来のボットは、リアリティのある実在の人間のようなペルソナには欠けており、定型的な投稿を繰り返すなどの行動から、一定のリテラシーがあれば見破ることは容易であった。しかし、ChatGPTなどのテキスト系生成AIサービスは、幅広いトピックにわたってリアリティのあるテキストを生成できるため、従来の無味乾燥なボットを強化する機会を提供してしまうこととなった。ある先行研究では、人間のようなコンテンツを生成するためにChatGPTを採用していると考えられる、Twitterのソーシャルボットネットワークが発見されている[6]。これは、特定のツイートにおいて、ChatGPTがコンテンツポリシーに違反するツイートを作成するように促された場合、自動的にChatGPTの謝罪メッセージ（(「有害または不適切なコンテンツの生成に関するOpenAIのコンテンツポリシーに違反するため、申し訳ありませんが、このリクエストに応じることはできません」)。）をボットがTwitterに投稿したことから明らかになったものである。現時点ではこのように看破できたボットであっても、今後、ChatGPTのようなガイドラインを有しないオープンソースの生成AIサービスを利用したり、この事例のような当該サービスの定型メッセージを出力したツイートをフィルタリングしたりすることで、より検出は困難となる。

さらには、影響工作のためのツールをパッケージしたソフトウェアをイスラエルの業者が販売していることがガーディアン紙のレポート[7]により明らかにされたが、こうしたソフトでも生成AIが利用されている。レポートの調査対象企業のひとつが、「Advanced Impact Media Solutions」（Aims）と呼ばれるソフトウェアを提供しており、このソフトウェアを使えば、Twitter、LinkedIn、Facebook、Telegram、Gmail、Instagram、Youtube上の何千ものフェイクアカウントをコントロールできるのである。

こうした生成AIによるコンテンツ生成は、リアリティあるディスインフォメーションの大量生成をも可能にする。ある先行研究[8]における実験では、公開されているある生成AIのモデルを使用すると、65分以内に、英語で17 000語以上のディスインフォメーションを含む102の異なる医療系ブログ記事を生成することが出来た。これらの記事には、偽の患者や臨床医の証言が含まれ、科学的に見える参考文献までもが含まれていたという。また、画像の場合は、2分間で20枚のディープフェイクを作成することが可能であった。こうしたディスインフォメーションおよびディープフェイクのコンテンツ生成は、中国やロシア、イランの影響工作において活用されるようになってきている[9]。

（2）ディープフェイクの脅威

生成AIによるもう一つの大きな脅威が、「ディープフェイク」である。ディープフェイクとは、「ディープラーニング」と「フェイク」を組み合わせた造語で、広義には、AIや機械学習によって生成・編集されたメディアやそのための技術のことである[10]。狭義には、人をだます目的で、写真、音声、映像の一部を入れ替えて、本物そっくりに合成された偽画像、偽音声、偽映像を指す。例えば、ある人のスピーチ動画で、その顔を別人の顔画像と入れ替えたり、別の音声と入れ替えたりするような操作がされたものがディープフェイクである。

サイバーセキュリティ会社センシティの調査[11]によると、2017年にインターネット上でAIによるディープフェイク動画が確認されてから、その数は急速に増加している。2018年末には約8,000件だったものが、2020年末には約85,000件にまで増加した。当初はポルノ動画作成が主要な目的であったが、

[6] Yang, K. C. & Menczer, F., “Anatomy of an AI-powered malicious social botnet. Observatory on Social Media,” 2023, Indiana University. ArXiv. (https://arxiv.org/pdf/2307.16336.pdf.)
[7] Kirchgaessner, S., et al., “Revealed: the hacking and disinformation team meddling in elections,” 2023, The Guardian. (https://www.theguardian.com/world/2023/feb/15/revealed-disinformation-team-jorge-claim-meddling-electionstal-hanan)
[8] Menz BD, Modi ND, Sorich MJ, Hopkins AM. Health Disinformation Use Case Highlighting the Urgent Need for Artificial Intelligence Vigilance: Weapons of Mass Disinformation. JAMA Intern Med. 2024;184(1):92–96. doi:10.1001/jamainternmed.2023.5947
[9] OpenAI, “AI and Covert Influence Operations:Latest Trends,” May 2024. (https://downloads.ctfassets.net/kftzwdyauwt9/5IMxzTmUclSOAcWUXbkVrK/3cfab518e6b10789ab8843bcca18b633/Threat_Intel_Report.pdf)
[10] 笹原和俊『ディープフェイクの衝撃　AI技術がもたらす破壊と創造』PHP新書、2023年、46頁。
[11] SENSITY TEAM, “How to Detect a Deepfake Online: Image Forensics and Analysis of Deepfake Videos,” 8 FEBRUARY 2021, sensity. (https://sensity.ai/blog/how-to-detect-a-deepfake/)