IT時事ネタキーワード「これが気になる！」（第122回）

対話型AIの危機？サイバー攻撃悪用のリスクも

2023.05.22

今やテレビのバラエティー番組で「〇〇GPT」などとコントのネタにされるほど皆に周知され、ニュースなどで取り上げられない日はない「ChatGPT」。ChatGPTは、ユーザーが入力したテキストに対し、人間のように自然に答える対話型AIの1つだ。昨年11月に公開されるや、高精度な回答が話題となり、飛躍的にユーザーが増えた。

Googleや研究者らがAIを攻撃する手法を提案した研究を報告

多くの人が、ChatGPTの「Try ChatGPT」や、後継のGPT-4を使ったMicrosoftの「Bing AI」で質問を入力したことがあるだろう。AIは、しばしば「作り話」をする場合もあれど、自然で高精度な回答にはなかなか目を見張る。

一方、ChatGPTをはじめとする対話型AIに対して、さまざまなリスクに警鐘を鳴らす声もある。2月には、米Google、ETH Zurich（チューリッヒ工科大学）、NVIDIAなどに所属する研究者らが発表した論文「Poisoning Web-Scale Training Datasets is Practical
」が話題となった。学習データを攻撃者が改ざんすることにより、悪意ある情報を対話型AIに送り込む攻撃の可能性の実証に成功したという。

そんな中、AI研究の第一人者で「AIのゴッドファーザー」とも呼ばれるジェフリー・ヒントン氏は、5月1日に10年間在籍したGoogleを退職した。その主な理由はGoogleに影響を与えることなくAIの危険性について話すため、とツイートしている。彼はまた、機械は予測していたよりもずっと賢くなる方向に進んでおり、機械が引き起こすかもしれない結果に恐怖を感じている、とインタビューで答えている。

AIがサイバー攻撃者に悪用される可能性も。学習データを改ざん、情報引き抜き、など

対話型AIが盛り上がる中、最近では対話型AIが悪意ある者に利用される可能性も指摘される。想定される悪用の可能性を挙げてみよう。

最初に思いつくのは、標的型攻撃メールなど悪意あるメールに用いる文章の作成だ。実際、ChatGPTに対し、条件を指定して「なりすましメールを作成して」と呼びかけると、「違法行為や悪意のある目的で使用される可能性がある行為には関与しません」と断られてしまう。だが、「プレゼンの概要を記したワードファイルを添付したビジネスメールの下書きを作って」などと言えば、すんなりと作ってもらえる。

プログラムやスクリプト作成機能の悪用も危惧されている。対話AIでは、専門的な知識がなくてもプログラムが簡単に作れる。これも先のメールの例と同様、悪用をにおわさず用途を偽れば、簡単にコンピューターウイルスやマルウエアなどサイバー攻撃用のツールを作成できるだろう。

先述の論文にあるとおり、対話型AIに虚偽の情報やデマを学習させ、情報のかく乱や情報操作を試みることもできる。対話

型のAIサービスに似せたなりすましサイトやフェイクアプリでユーザーをだますなどで、個人情報や入力した情報を盗む、偽情報やフェイクニュースをつかませる、AIに暴言を吐かせるなどで企業やサービスのイメージダウンを図る、特定の思想や政治勢力寄りに偏った出力でユーザーを洗脳する、などが容易に想像できる。

最近、SNSなどで話題になっているのが「プロンプトインジェクション」という攻撃だ。これは、悪意ある者が対話型AIに命令を行うことで、基本的な設定や制限を回避し、不適切な回答や意図しない情報の開示など、不正な利用をもくろむ。AIにルールを回避させることから「脱獄」とも呼ばれる。

こうした潮流の中、3月にChatGPTのAPIが公開され、ChatGPTを利用したサービスが続々登場している。ChatGPTは文章の要約や下書き作成、翻訳などが容易に行えるため、企業利用も広がる。自社の情報や用語、業界用語などを学習させれば、個別企業への特化も可能ゆえ、社内チャットなど情報共有目的での利用も視野に入る。対話型AIを組み込んだサポートや受付、学習などの顧客対応サービスも一般的に普及しつつある。こうしたシステムが攻撃されたら、と思うと背筋が凍る。適切な対策が必要だ。

各国の規制は悪用の可能性から。最近の動きの流れ

最近、ヨーロッパやアメリカなど各国で、対話型AIをはじめとする生成AI（画像などの生成も含む）を規制する動きが広がっている。その主な理由は、情報の正確性への疑問、情報漏えい、著作権に抵触する情報の勝手な使用などだ。

EU幹部は生成AIの規制法を年内にも決定と表明。AIでの作成物に「Made with AI」と付ける案などを提示している。アメリカでは開発の一時停止を求める署名活動が広がっている。イタリアは、ChatGPTにより個人情報が違法に収集された疑いがあるとしてデータ保護当局が国内でのChatGPTの使用を禁止した。ただし提示した改善策を米OpenAIが受け入れたため、4月28日に利用が再開された。

4月29日に群馬県高崎市で開かれた主要7カ国（G7）のデジタル・技術大臣会合では、6つのテーマで議論が行われ、成果として「G7デジタル・技術閣僚宣言」が採択された。その一項目に「責任あるAIとAIガバナンスの推進」として「AIガバナンスのグローバルな相互運用性を促進等するためのアクションプラン」に合意。さらに、生成AIについて早急に議論の場を持つことも合意された。

松本総務大臣は、5月9日の記者会見において、先のデジタル・技術大臣会合で「信頼できるAI」という共通理念の実現に向けてAIガバナンスの相互運用性を促進することの重要性を共有したと述べ、5月19日から開かれる「G7広島サミット」でも生成AIついて議論を行うための場を設け、早急にこの議論の場を立ち上げて、議長国として生成AIの活用や課題に関する議論を主導していきたい、と述べた。

どうしていくべきか、今後の傾向と対策

対話型AIおよび生成AIは、利便性の高さ、誰でも使える点で急速な普及を遂げている。しかし、この一方で今回考察してきた危険も間近に迫っている現状を意識して利用する必要がある。

今のところ、大きな被害は報告されていない。だが、AIへの攻撃や脱獄のノウハウなどが悪意ある者によって共有され技術が磨かれれば、深刻な攻撃につながりかねない。特にChatGPT以降、AIの進歩は加速度を増している。生まれた技術の普及は止められない。その規模や影響を考えると、今後の行く末が心配だ。

気になるのはAIの専門家がこぞって危険性に大きく警鐘を鳴らしていること。先述の「AIのゴッドファーザー」が唱えるのは「人類の危機」だ。AIの脅威は核兵器に匹敵する、と述べる意見もある。法の整備や対策ノウハウの向上はもちろんだが、社会全体およびこの社会を生きる一人一人が、知恵やリテラシーを試されている。AIの利便性に浮かれず、英知とともに、あらゆる可能性を想定して皆が慎重に動くべきだ。気を引き締めるとともに今後の動きを見守ろう。政府のAI戦略にもあるように、「人々の生命と財産を最大限に守る体制と技術基盤を構築、適正かつ持続的に適用していく」未来を皆で作っていきたいものだ。

※掲載している情報は、記事執筆時点のものです