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… 続きを読む