科学技術の潮流 ⑳

JST研究開発戦略センター

AIセキュリティー強化を

AI（人工知能）技術の発展が目覚ましく、高度な処理の自動化による生産性向上、新たな事業機会の可能性に期待が高まっている。一方で、AIの脆弱（ぜいじゃく）性や悪用・誤用に起因するリスクが深刻化し、安全性確保のためのAIセキュリティー技術が不可欠になった。

安全性に焦点

AIについて、安全性を含む社会的要請が国際的に議論されるようになったのは2019年である。日本政府は「人間中心のAI社会原則」を発表し、経済協力開発機構（OECD）がまとめたAI原則に42カ国が署名した。公平性、透明性、説明責任、プライバシーなど、AIに倫理性を求めた。

22年に人間と区別困難な応答を返す生成AIが登場し、偽・誤情報の生成・拡散による世論誘導や犯罪の巧妙化といった悪用問題、生成AIを誤動作させる攻撃法など、AIの安全面の懸念が増大した。その結果、安全性が強く要請されるようになり、23年に英国主催でAI安全サミットが開催されたのを契機に、安全性に関する評価手法や基準の検討・推進を行う機関「AIセーフティ・インスティテュート」が英米日ほかで設立された。

原則から実践へ

シアムから19年に、産業技術総合研究所から20年に、それぞれガイドラインの初版が公開され、開発現場での活用も広がっている。24年には経済産業省と総務省による「AI事業者ガイドライン」も公開された。これらは開発者・事業者に適正な取り組みを促し、利用者の意識を高めるのに役立つ。

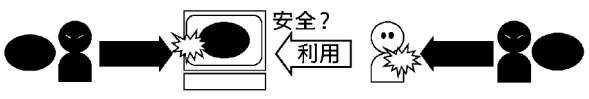
しかしそれだけでなく、AI特有の脆弱性を突いて誤動作させたり情報搾取したりするAIセキュリティー技術の開発が急がれている。

AIはさらに発展し、自ら計画を立てて実行したり、他のAIと連携・交渉したりするようになりつつある。そこで生じる新たなリスクに先んじて対処するためにも、AIセキュリティー技術の一層の強化が急務である。**（金曜日に掲載）**

科学技術振興機構（JST）研究開発戦略センター
フェロー（システム・情報科学技術ユニット）
福島 俊一

東京大学理学部物理学科卒、IT企業にて自然言語処理・情報検索の研究開発に従事後、16年から現職。工学博士。11―13年東大大学院情報理工学研究科客員教授、情報処理学会フェロー。

「AI利用・運用時のリスク」



(a)AI特有の脆弱性を突いて誤動作させたり情報搾取したりする

(c)安全なAIだけでなく粗悪なAIや邪悪なAIが潜み利用者を欺く

(b)フェイク生成等AIを使って人をだましたり誘導したりする

［出典］研究開発戦略センター（CRDS）「人工知能研究の新潮流2025～基盤モデル・生成AIのインパクトと課題～」を基に筆者作成

無断転載・複写禁止 ㈱日刊工業新聞社