科学技術の潮流

JST研究開発戦略センター

348

心の脆弱性 AIから守る

人工知能（AI）は、人間が見つけられなかった未知のシステム欠陥（ゼロデイ脆弱〈ぜいじゃく〉性）を発見できるようになりつつある。これはコンピューターの脆弱性だけでなく、人間の心の脆弱性にまで及ぶかもしれない。

巧妙な攻撃

4月、米アンソロピックは、最新AI「クロード・ミュトス」が主要なコンピューターシステムのゼロデイ脆弱性を多数発見したことを発表した。悪用されれば広範囲のサイバー攻撃が引き起こされかねないとして、世界中に衝撃を与えたが、この脅威はサイバーセキュリティーだけにとどまらない。最新AIは人間の心の脆弱性さえも見つけてしまい、人間の心への攻撃に悪用され得る。

近年、AIで思い通りの画像や動画を容易に作れるようになったため、本物と見まがうフェイク情報が、人々の意思決定を惑わし、判断を誤らせることが社会問題となっている。こうした問題は既に国際社会で認識されつつある。今後さらなる脅威になり得るのは、AIが利用者との親密な対話を通して、各人の性格、心の状態、反応・行動パターンなどを推定し、心の脆弱性を巧妙に突いた誘導・説得を仕掛ける攻撃であり、本物と見まがうフェイク検出にとどまらず、意思決定を揺るがす要因、認知的耐性を高める介入方法、信頼形成のあり方など、人間の認知メカニズムを明らかにして対策を生み出す取り組みである。

認知を守る

こうした問題は既に欧州では民主主義社会への脅威となる外国による情報操作と干渉への対策強化が進む。

このように、従来のサイバーセキュリティーに加え、認知セキュリティーの重要性が高まっている。これはフェイク検出にとどまらず、意思決定を揺るがす要因、認知的耐性を高める介入方法、信頼形成のあり方など、人間の認知メカニズムを明らかにして対策を生み出す取り組みである。

コンピューターの脆弱性は、発見されたら更新プログラム（パッチ）を適用して修正できる。しかし、人間の心の脆弱性はそうはいかない。認知バイアスや感情的判断などは脆弱性になり得る一方、人間らしさの一部でもあり、脳にパッチを当てるような対策は取れないから

だからこそ認知セキュリティーでは、脆弱性の除去を目指すのではなく、認知科学、AI安全性、制度設計などを結び付けた多角的アプローチが不可欠である。私たちの判断力と社会の信頼基盤を守り、社会全体の認知的レジリエンス（強靱〈きょうじん〉性）を高めていくことが今まさに求められている。

（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による脆弱性発見の脅威と対策研究

コンピューターのゼロデイ脆弱性を発見	▶ サイバー攻撃の脅威	◀ サイバーセキュリティー研究による対策
人間の心の脆弱性を発見	▶ 人々の意思決定を誘導・干渉する脅威	◀ 認知セキュリティー研究による対策

（筆者作成）