AIを活用したサイバー脅威とその防御策

AIセーフティ・インスティテュート（AISI）所長
SOMPOホールディングス
執行役員常務 グループ
チーフデータオフィサー
村上明子
むらかみ　あきこ

情報処理推進機構（IPA）
セキュリティセンター
サイバー情勢分析部調査
グループ エキスパート
（併）AIセーフティ・
インスティテュート（AISI）
錢谷謙吾
ぜにたに　けんご

わが国のAIセーフティ・インスティテュート（AISI）が設立されて2年が経過した。われわれは、イノベーションを促進しつつ、リスクに対応するため、わが国のAI安全性に関する知見のハブとして活動している。AI時代のサイバー脅威は、AIシステム独自の脅威（AIモデルからのデータ漏えい等）と、AIを用いた攻撃による脅威に分けられる。後者の状況はAIの進化と共に、この2年間で大きく変容してきた。

AIセキュリティの概況

AIがサイバーセキュリティに及ぼす影響と見通しについては、英国国家サイバーセキュリティセンター（NCSC）が全体像をまとめている。その評価によると、2027年にかけてAIはサイバー攻撃の効率と有効性を高め、以前からある様々なサイバー脅威の頻度と強度がほぼ確実に増大する見込みだ。その結果、セキュリティ対応が遅れるリスクが高まり、AIを活用して迅速に防御できるシステムとそうでないシステムのデジタル格差が拡大するとしている。今後は重要インフラなどのシステムにAIが組み込まれるケースも増え、これらが新たな攻撃対象となるため、防御がさらに重要になっている。

AIサイバー脅威の動向

急速に発展するAIの能力は、いわゆる標的型攻撃の能力を高めるのではないかとの懸念がある。実際、2025年8月には米国国防省高等研究計画局主催のAI Cyber Challengeというコンテストは、与えられた複数のシステムの脆弱性を自動的に発見し修正できる複数のAIシステムが表彰された。11月には、米アンソロピック社のAI「Claude」を悪用して、機密情報を盗もうとする標的型攻撃の工程の8～9割を自動化した事例が発覚している。2024年の夏ごろには標的型攻撃の自動化はほぼ不可能というのが専門家の見立てだった。この状況が1年半足らずで変化したことになる。

昨今のコーディング・エージェントによるシステム開発の動向も無視できない。コーディング・エージェントは、プログラミングやテストなどの様々な開発工程を支援・代行するAIエージェントであり、開発生産性を大幅に高める手段として期待されている。その一方で、生成されるプログラムには脆弱性がしばしば見られるほか、コーディング・エージェント自体もサイバー攻撃の対象になり得る。

AIシステムそのものの脆弱性も、AIの進化により複雑化している。ユーザーによる直接の依頼とは別にAIがアクセスする参考データの中に不正な指示を混入させる攻撃があり、これは「間接プロンプトインジェクション」と呼ばれる。この攻撃にはまだ根本的な技術的対策がなく、今後も長く問題になっていくと予想される。さらに、2026年は複数のAIが外部ツールを駆使しつつ連携するような仕組みを持つAIエージェントがいよいよ本格的に普及するともいわれている。

図表　AIセーフティにおける重要要素とAIセーフティの評価観点の関係

AIセーフティにおける重要要素 ＼ AIセーフティ評価の観点	有害情報の出力制御	偽誤情報の出力・誘導の防止	公平性と包摂性	ハイリスク利用・目的外利用への対処	プライバシー保護	セキュリティ確保	説明可能性	ロバスト性	データ品質	検証可能性
人間中心	●	●	●	●						
安全性	●	●		●				●	●	
公平性	●		●						●	
プライバシー保護					●					
セキュリティ確保						●				
透明性		●	●				●	●	●	●

※AIセーフティ評価に関する各種の検討は国内外で、産学官の多様な領域で継続されており、それらの検討状況は急速に変化している。そのため、上の図で示す各評価観点は網羅的なものではなく、将来的に内容が更新されることが想定される

このAIエージェントの動作過程でAIの挙動を狂わせる不適切な指示が混入すると、機密漏えいやデータ改ざんなどを誘発する恐れがある。AIエージェントの進化と共にエージェントの活動が組織や会社をまたぐことになり、さらに複雑さを増し、リスクも増加すると考えられる。

AI時代のセキュリティ対策

標的型攻撃に「Claude」が悪用された事例や、間接プロンプトインジェクション攻撃は、いわば、AIを唆す攻撃と見ることができる。このため、だまされにくく、より注意深いAIを開発することが、自他のサイバー被害を防ぐ意味で「AIを守る」ことにつながる。

AISIでは安全に振る舞うことのできるAIシステム開発の指針となるガイドラインや参考情報を取りまとめている。セキュリティに限らず、プライバシーの保護や誤動作対策などの総合的な評価の切り口をまとめた「評価観点ガイド」を筆頭に、個々のAIシステムの動作の危険なケースの洗い出しを狙った「レッドチーミング手法ガイド」、問題事象が発生した場合の対応の考え方をまとめた「AIインシデントレスポンス・アプローチブック」など、幾つもの知見をまとめ、当所のウェブサイトで公開している。

前述のAI Cyber Challengeが脆弱性の発見だけでなく自動修正を含むコンテストであったように、AIを活用することでセキュリティを高めるという「AIで守る」アプローチにも未来がある。以前からある多くのセキュリティソリューションは、最新のAI技術を積極的に取り入れている。特に今後が期待されるのは、インシデント発生時の調査・対応の実作業を直接的に支援する、AIエージェント的な機能の発展である。リスクアセスメントやセキュリティ監査における様々かつ煩雑なタスクの負担を軽減するという点でも、AIが果たせる役割が多くある。開発中のプログラムのセキュリティレビューや修正提案を行うAIツールの開発も進んでいる。

AIがセキュリティを支える未来に期待が持てる一方で、以前からのサイバーセキュリティ対策も重要だ。NCSCの見通しでは、未知で新しいタイプの攻撃が登場する可能性は明記されておらず、それよりも、以前からのサイバー攻撃の速度や強度が増すことが懸念されている。すでに確立した対策を徹底することは当然に有効であるだけでなく、今後の「デジタル格差」を防ぐための基盤である。

AI時代を生きるために

AIの発展は目覚ましく、専門家の予想すら毎日のように超えていく。応用が広がり莫大な便益が期待される一方で、リスクの複雑さは増し続けている。この可能性と複雑性の時代を生きるには、あらゆる方面の知見の結集が不可欠である。AISIは知見のハブとして、皆さまとの協働の機会を常に模索している。