# AIセキュリティ短信 2025年9月号

※注意※

本稿はAIセキュリティに関連するニュースを収集・選別し要約したものです。個々の事例の記載内容は参照している情報源の記載に準じたものであり、その内容の正確性・妥当性等をIPAにおいて保証するものではなく、また、参照される製品・ソリューション等を批判・推奨するものでもありません。

## 今月の話題

① サイバー防御の自動化を目指すDARPA AIxCCが決着
② 脆弱性検知ランキングでAIが人間を超え1位獲得
③ LLMを使ったマルウェアがウクライナで発見
④ 生成AIを利用したサプライチェーン攻撃が登場
⑤ 米トランプ政権がAIアクションプランを発表
⑥ 間接プロンプトインジェクションの事例が続出

## ①サイバー防御の自動化を目指すDARPA AIxCCが決着

AI Cyber Challenge（AIxCC）は、国防高等研究計画局（DARPA）と保健医療高等研究計画局（ARPA-H）が主導し、2年間で総額2950万ドルの賞金がかけられたコンペティションである。AIxCCは、重要インフラを支えるオープンソースソフトウェアに潜む脆弱性を「自律的に発見し、修正する」革新的なAIシステムを創出することとされた。最終結果は2025年のDEF CON 33で発表され、ファイナリストチームは、主催者が意図的に埋め込んだ70件の「合成」脆弱性のうち54件と、課題とは別に、これまで未知であった18件のゼロデイ脆弱性も発見した。脆弱性のパッチ生成と適用にかかる時間は平均わずか45分、脆弱性1件あたりの修正コストは平均152ドルと試算されており、DARPAのプログラムマネージャーであるAndrew Carney氏は、セキュリティ自動化の「新たな最低基準（the new floor）」を確立したと宣言した。

## ②脆弱性検知ランキングでAIが人間を超え1位獲得

2025年6月、AI駆動型の自動ペンテスターであるXBOWは、HackerOneの米国リーダーボードで第1位にランクインというAI史上初の快挙を成し遂げた。HackerOneは世界最大級のバグバウンティプラットフォームである。XBOWは過去90日間で約1,060件の脆弱性を提出し、その影響はAT&T、Ford、The Walt Disney Companyといった名だたる企業に及び、現実に運用されているシステムの脆弱性検知における実力を示した。XBOWは、単に脆弱性の可能性を指摘するのではなく、脆弱性の種類に応じて検証プログラムをカスタム構築し、その存在を機械的に検証する仕組みを導入し、検知精度を大幅に高めた。ただし、機械的な検証が可能な脆弱性を扱うという性質から、検知できるのは中～低深刻度の脆弱性が大半であって、高度なゼロデイ脆弱性が容易に自動検知可能になったわけではない。

## ③LLMを使ったマルウェアがウクライナで発見

2025年7月、ウクライナで発見されたマルウェアLameHugは、侵入対象となった環境の詳細情報を収集し外部送信する攻撃準備段階を担うものであったが、その環境調査処理を生成AIにその場で自動プログラミングさせていたことで大きく注目を浴びた。翌8月には、PromptLockと呼ばれるマルウェアも確認された。PromptLockはランサムウェア攻撃の一環となる、暗号化すべきファイルの探索・選定・暗号化といった複数のタスクのためのコード生成をAIに委ねており、LameHugよりも一歩進んだ生成AIの悪用を行っている。AIによるコード生成結果は毎回異なるため、シグネチャベースの検知手法では見落とされることとなり、生成AIを用いた検知困難なマルウェアの事例として今後の長期的な警戒を喚起するものと言える。

## ④生成AIを利用したサプライチェーン攻撃が登場

2025年8月26日、サーバサイドJavaScriptアプリ開発で広く用いられているビルドツールNxの脆弱性をついたサイバー攻撃が発見された。そもそもの起点となったNxの脆弱性がAI生成コードに含まれる欠陥であったということに加え、混入されたInfostealerが生成AIを利用して窃取対象のデータを探索するスクリプトを生成しているという2点で、本事例は生成AIとの注目すべき接点を有する。この攻撃により400以上の組織が被害を受け、5,500以上のプライベートリポジトリが公開されるという

二次被害が確認された。Nx はアプリケーション開発の基盤となるツールでもあり、生成 AI を利用した初のサプライチェーン攻撃の確認事例ともなっている。

## ⑤米トランプ政権が AI アクションプランを発表

2025 年 7 月 23 日、米トランプ政権は America's AI Action Plan を発表した。このアクションプランは前バイデン政権の大統領令 EO 14110（通称 AI 大統領令）を置き換える新しい大統領令 EO 14179 に基づき、EO 14179 の発布から 180 日以内に策定するよう求められていたものである。本 AI アクションプランはアメリカの AI 覇権（AI dominance）確保が明確に打ち出されており、セキュリティについての言及も重要インフラの保護と結びついた国家安全保障色の強いものとなっている。アクションプランに基づく具体的なガイドライン等の整備はこれからの動きとなるが、AI ISAC の設置、AI Incident Response の準備などが謳われており、今後の動向が注目される。

## ⑥間接プロンプトインジェクションの事例が続出

EchoLeak 以降、信頼できないデータに不正プロンプトを埋め込む間接プロンプトインジェクションが多様化している。Web ページ中の隠しテキストで AI ブラウザ中の LLM を唆し情報を盗む事例や、縮小すると命令が浮かぶ画像を悪用する事例に加え、カレンダーの招待状経由でスマートホーム機器を不正操作し、物理世界にまで影響を及ぼす攻撃も確認された。LLM が入力中の不正プロンプトを確実には識別できない根本問題への特効薬はまだなく、AI に与える権限の最小化や、重要な操作にはユーザーの許可を必須とする仕組みの導入など、基本的なセキュリティ対策の徹底がこれまで以上に重要となっている。

## 編集後記

今号では、8 月に DEF CON 33 が開催されたことを念頭に、AI-Enhanced Threat という切り口で注目される事例を中心に紹介いたしました。AI によるサイバー攻撃の自動化も LLM 搭載型マルウェアもどちらも現実化している一方で、その内実は従来のサイバー攻撃やマルウェアを効率化・高度化したものに当たり、これまでにない全く新しいタイプの脅威が登場したとまでは言えない状況です。AI セキュリティの概況には英 NCSC が 2024 年 1 月と 2025 年 5 月に見通しを示しており、警戒は必要であるも

のの、AIの進歩がサイバーセキュリティに及ぼす影響は、深刻ではあっても想定外とまでは言えないというある種の相場観が形成されつつあるやに思われます。

他方、今号では意見交換会に先立ちドラフト版を皆様にお届けした上で、ご意見・フィードバックを頂戴いたしました。今号⑥の間接プロンプトインジェクションはいただいた情報を元に追記した内容となります。ご意見をいただいた皆様には改めて厚く御礼申し上げます。まだまだ手探りで進め方も定まりませんが、今後ともご助言・ご協力をいただけますと幸いです。