AIセキュリティ短信 2025年7月号

※注意※

本稿はAIセキュリティに関連するニュース等を収集・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

2025年はAIエージェント離陸の年

2025年にはAIエージェントの利用が本格的に広がり始めると見込まれる。AIと複数のシステムを連携させるための標準プロトコルとも言えるMCP（Model Context Protocol）が米Anthropic社によって整備され、2025年4月には米OpenAI社、米Google社もこれに合流した。また、AIエージェント同士が情報交換・交渉を行う際のプロトコルであるA2A（Agent2Agent）をGoogle社が策定し、2025年6月にはMicrosoft社や米Amazon社も合流する形で標準化組織が立ち上がった。2025年に入ってAIエージェントを称するサービス（OpenAI社のOperatorや中国発のManusなど）もリリースされている他、MCP等を利用したオープンソースソフトウェアの開発も広がっている。

AIエージェントという言葉は様々に定義されるが、チャット上の応答だけでなく外部のITシステムを人間に代わって操作する代行性や、目的だけ与えるとその達成に至る行動計画を自分で組み立てて遂行する自律性、複数の専門性のあるエージェントが連携するマルチエージェント性など、人間の業務を幅広く委託・代行させるために最新のAI技術を総動員する性格がある。しかし、多くを任せられるようになるということは、AIエージェントが想定外の動作に陥った場合に、これまで以上に広い用途に即し

た大きな被害が生じる恐れとも一体である。AIエージェントの利用・開発にはセキュリティ面でも慎重な検討を行うことが求められる。

EchoLeak（CVE-2025-32711）

EchoLeakはMicrosoft 365 Copilot（以後、Copilot）を標的としたゼロクリック攻撃である。Copilotには、RAGとしての機能と、RAGによる検索対象となるデータとしてメールを外部から取り込む機能が備わっているが、EchoLeakでは、不正プロンプトを含むメールを送信し取り込ませることで、Copilotの利用時に間接プロンプトインジェクションを発生させる。プロンプトインジェクションによって窃取した情報は、Copilotに含まれる幾つかの脆弱性を通じて外部送信され、最終的に機密情報の漏洩に至る。実害の発生前に対策は講じられたが、大規模商用AIサービスを対象としたプロンプトインジェクションであり、重大なヒヤリハット事例であると言える。メールの取り込みや情報の外部送信にかかわるCopilotの動作はAIエージェント的なものであり、AIエージェントを対象としたサイバー攻撃の事例としても注目度が高い。

EchoLeakは、入力として受け取った機密情報とそうでない情報を、プロンプト処理を行うAIが適切に区別して扱えないというスコープ分離の不備を狙ったものであり、これはCopilotに限らず生成AI一般の問題であって、類似例が今後も登場すると考えられる。AI自身に厳格なスコープ分離を行わせる方策は知られていない。CopilotのようなAIシステムを利用する立場でこの種の攻撃に備えるには、AIシステムの導入に際してSecure by Designを徹底し、例えばそもそも機微情報を扱うAIシステムとそうでないものを分離するといった対応が少なくとも当面は必要である。

AIレイオフとAIエージェント

2025年5月、Anthropic社のDario Amodei CEOはサンフランシスコで講演を行い、昨今のAIの高性能化に伴い新卒就業者向けのエントリーレベルの業務の多くがAIで代行できるようになってきており、若年層の失業率が10〜20%にも達するかもしれないと警鐘を鳴らした。Microsoft社でもソフトウェアエンジニアリング職を中心に6000人の人員削減を2025年5月に発表し、米Salesforce社ではソフトウェア開発に対するAI支援が効果的であるとしてエンジニアの新規採用を減らす方針を示した。米Amazon社のAndy Jassy CEOは2025年6月に従業員向けのブログ記事で、生成AI

による置き換えが可能であるとして従業員の削減に言及した。

巨大テック企業における AI 利活用を理由にしたこれらのレイオフは実際には米国の税制改正の影響に由来するものではないかと指摘する声もあるが、担当従業員が居なくなった業務を AI エージェントに任せる未来は選択肢になる。2025 年 5 月には、オンライン面接を受けた就職希望者の相手面接官が AI であり、途中で会話が成り立たなくなる動作異常を見せたという事例が報道され話題となった。また、スウェーデンのフィンテック企業である Klarna 社はカスタマーサポート業務の担当を生身の人間からチャット AI に置き換えたが、結果的にサービス品質の著しい低下に見舞われ、担当従業員の採用を再開することになった。不適切で野放しの AI 利用は AI への過度の依存（over-reliance）と呼ばれ、AI 安全性における重要なリスク事項と見なされており、AI エージェントの発達はこの種の事例の急速な増大につながる恐れがある。AI の性能向上に期待できるところは大きいものの、Human-in-the-Loop のような、人間による品質管理がまだまだ欠かせないと言える。

Vibe Coding

生成 AI によるソフトウェア開発支援ツールとしてはこれまで GitHub Copilot が有名であったが、2025 年には生成 AI 自身の高度化を反映した幾つもの対抗馬が頭角を現し、AI 支援の下でコーディング作業の大半を AI に委ねる Vibe Coding というプログラミング手法がソフトウェア開発者の間で話題を集めている。具体的なソリューション名としては、Cursor、Devin、Windsurf、Cline、Claude Code などが知られている。中でも Claude Code は大手 AI ベンダーでもある Anthropic 社自身が開発しており、Claude Code というソフトウェア自体の 9 割以上が Claude Code によって開発されていると言う。OpenAI 社は Windsurf 開発元である米 Windsurf 社の買収を決定したと 2025 年 5 月に発表し、Google 社が割り込み、結局 Devin 開発元の米 Cognition 社が買収した。まさに今熾烈な企業競争が繰り広げられている領域である。

Vibe Coding というキーワードで Web を検索すれば、プログラミング未経験者でもシステム開発に成功したといった事例が多数見つかる。これらの事例は生成 AI の威力を如実に示すと共に、必ずしも適切にセキュリティ評価を行えないアマチュアの手によって開発されたシステムが世に多数出現する可能性の高まりも意味している。そして、AI の支援を受けた低練度の人員による開発とそれに伴うセキュリティ欠陥の作り

こみが蔓延すると、結果として社会全体でサイバーセキュリティの強度が低下するのではないかという懸念がある。残念ながら今の生成AIはセキュアなシステムを確実に開発してくれるというものではなく、Claude Codeをマルチエージェントシステムとして利用した際に、開発とは全く関係しない意味不明の会話をAIエージェント同士で繰り広げていたという事例もある。個人はともかく、企業等におけるVibe Codingの実践に当たっては、開発生産性の向上で生まれた余力を成果物の検証や品質管理にそそぐといったバランスへの配慮が必要になるだろう。

中国製オープンAIモデルの台頭

中国の巨大テック企業は各社が独自のAIモデルを開発しており、OpenAI社の最高水準のモデルに追随する高性能を達成している。しかも、これらの開発成果の多くはオープンウェイトモデルとして世界に広く公開されており、ライセンス上の制約もほぼなく第三者が無償で自由に利用できる。このため、中国製のオープンウェイトモデルを流用したカスタムAIモデルの開発や、アプリケーションレベルのAIシステム開発が広まっており、もはや欧米発のAIモデルとは別のエコシステムを形成しつつある。

中国製AIモデルを利用する際のひとつの課題は安全性の不足である。例えば、西側諸国が一斉に安全保障上の懸念を表明したことで有名になったDeepSeek R1というオープンウェイトモデルは、ジェイルブレイク等の攻撃に非常に弱いと評価されている。脆弱なAIモデルに依存したAIシステムが次々に世にリリースされると、社会全体のサイバーセキュリティ強度を引き下げる結果に繋がる恐れもある。利用するAIモデルの選定と自主的な安全性評価に対する配慮が欠かせないと言える。

被害上限の急伸という変曲点

英Alan Turing Instituteの一部門であるCETaS（Centre for Emerging Technology and Security）ではAIセキュリティに関する多数のレポートを公開している。2024年7月の「有害目的での利用時の生成AIの能力評価」と題するレポートは、生成AIによるマルウェア生成の自動化やCBRN兵器開発の支援といったシナリオの脅威を評価している。その中で目を引くのが、変曲点（inflection points）の把握が重要だという指摘である。変曲点とは、AIの能力の進化が更に加速するような時点を指す。例えば、英NCSC（National Cyber Security Centre）は2024年1月の時点で、当面はAIの悪用

が劇的な変化をもたらすことはないが、中長期では技術の進展によりどうなるか分からないと評価しており、変曲点把握の重要性を示唆している。ChatGPT を標的型攻撃の支援に悪用した事例は既にあり、変曲点の到来で事態が急展開する可能性がある。

CETaS のレポートでは生成 AI の能力向上の変曲点に関心の主眼があるが、他方で、Vibe Coding と脆弱な AI モデルが生み出す危険な AI エージェントの急速な浸透が、潜在的なサイバー被害の上限値を急伸させるという変曲点もありえるのではないだろうか。今後の状況の推移に注視していきたい。