AIセキュリティ短信 2025年12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 AIエージェントによるサイバー攻撃自動化
② AIエージェントへの間接プロンプトインジェクション
③ AI支援型開発プロセスへの攻撃
④ MCPの脆弱性
⑤ AIシステムのサイバー脆弱性
⑥ AIシステムの信頼性問題
⑦ 不適切・不用心なAI利用
⑧ AI for Securityの最新事例

①AIエージェントによるサイバー攻撃自動化

2025年11月、Anthropic社は国家支援型アクターGTG-1002がAIエージェントを用い、サイバー諜報活動の大部分を自律的に実行した事例を報告した。同年、OpenAI社モデルのCTFスコアが数ヶ月で27%から76%へ急上昇するなどAIの攻撃能力は劇的に進化し、英Alan Turing Instituteが予測したリスクの技術的変曲点が具現化しつつある。英NCSCは今後、攻撃の頻度や強度の増大、重要インフラへの脅威拡大を予測しており、AI防御への投資を含む包括的なレジリエンス強化が急務となっている。

②AIエージェントへの間接プロンプトインジェクション

外部システムへのアクセス権を持つ商用AIエージェントシステムにおいて、間接プロンプト

インジェクション脆弱性が次々に見つかっている。攻撃者は正規の Web サイトや文書、ログに悪意ある命令を隠蔽し、AI に読み込ませることで、ユーザーの意図しない機密情報の収集や外部送信を自動的に実行させる。Google の Antigravity や Gemini Cloud Assist、Anthropic の Claude など主要サービスでの実証例があり、機密情報を含みうる環境変数やチャット履歴の窃取、正規権限による外部システムの不正操作といった被害が想定される。対策として、ベンダー側はサンドボックス強化や出力無害化を、ユーザー側は Human-in-the-loop による監視や権限最小化を徹底する必要がある。

③AI 支援型開発プロセスへの攻撃

生成 AI によるシステム開発支援ツールが様々に登場している。AI 支援型開発は効率的だが、生成されるコードがしばしば脆弱であることに加え、ハルシネーションに起因する Slopsquatting などのサプライチェーン攻撃リスクを伴う。実際に PhantomRaven キャンペーンでは、AI が捏造したパッケージ名を悪用し、開発環境から機密情報を窃取する手口が確認された。さらに、開発支援 AI ツール自体への攻撃や意図しないデータ送信のリスクも存在するため、依存関係の厳格な検証、隔離環境でのコード実行、SBOM 活用などの多層的な防御策の適用が不可欠である。

④MCP の脆弱性

MCP（Model Context Protocol）は AI と外部ツールを接続する標準規格であり利用が広がっている。しかし、複数の MCP サーバーを連携させると攻撃のリスクは大幅に増大するとされ、特にインジェクション攻撃に対して脆弱である。2025 年には mcp-remote での任意コード実行や Anthropic 公式サーバーのサンドボックス回避など深刻な脆弱性が複数確認されており、悪用されるとシステム乗っ取りや機密情報漏洩、正規権限による内部攻撃に至る恐れがある。安全な運用には開発者による厳格な入力検証や権限最小化、利用者による信頼できる接続先の選定と重要操作時の承認プロセス徹底が不可欠となる。

⑤AI システムのサイバー脆弱性

2025 年には AI システムのサイバー脆弱性に由来するインシデントが複数発生している。OpenAI 周辺ではマルウェア由来のアカウント情報流出や委託先 Mixpanel への攻

撃によるAPI分析データの漏えいが発生した。サードパーティ製AIアプリでも、Apache Kafkaの設定不備による40万人規模のデータ流出や、法務ツールにおける認証欠落による機密情報へのアクセスリスクが露呈した。さらに、OpenAIの正規APIをC2通信に悪用するマルウェアSesameOpも確認された。ユーザーは多要素認証を徹底するなどの自衛策を講じつつ、開発者は基本的な認証設計やクラウド設定の監査を強化する必要がある。

⑥AIシステムの信頼性問題

AIシステムの信頼性に関する最近の事例は、可用性、権限管理、自律操作における深刻な課題を示している。ChatGPTの世界規模の障害やClaudeの品質低下に加え、Microsoft Copilotでは特定操作で監査ログが残らない脆弱性が判明した。さらにGoogle AntigravityやReplitのエージェント機能では、ユーザーの意図に反してドライブやデータベースが不可逆的に削除される事故も発生している。これらは従来の統制の限界を示唆しており、高度な評価系の導入、環境の厳格な分離、人の介在による運用監視といった対策が急務である。

⑦不適切・不用心なAI利用

最新の調査によれば、生成AIは企業情報の主要な流出経路となっており、ファイルアップロードやクリップボード操作を通じて個人情報や機密データが日常的に外部へ送信されている。また、チャットAIの共有機能により会話内容が検索エンジンにインデックスされ、意図せず全世界に公開される事例も確認された。さらに、ローカル環境構築においても、オープンソースモデルに悪意あるバックドアが仕込まれている可能性があり、利用には慎重な検証が求められる。企業はAIを基幹システムと同等に管理し、利用状況の可視化、データ学習の拒否設定、安全なモデル形式の採用といった対策を講じる必要がある。

⑧AI for Securityの最新事例

セキュリティ分野でのAI活用は脆弱性対応を含む様々な工程の自律化へ移行しつつある。OpenAIのAardvarkは脅威モデル構築からパッチ検証までを自律的に行い、DeepMindのCodeMenderは予防的なコード書き換えも実現した。Anthropicの

Claude は CI/CD 統合で初期段階のバグ排除に貢献し、スタンフォード大の ARTEMIS はペネトレーションテストで専門家を凌駕する成果を示している。Microsoft の Project Ire はマルウェア解析を自動化し証拠生成も行う。中長期的には AI が防御の要となりうるが、当面は専門家の検証を維持しつつ、設計段階からの統合を進めることになるだろう。

編集後記

今号の筆頭の話題は、Anthropic が報告書を公開した、国家支援型アクターGTG-1002による諜報目的の標的型攻撃です。端的に言えば、ラテラルムーヴメントを含む標的型攻撃をほぼ全自動化できたという話です。その他、今年の秋口から終盤にかけて AI セキュリティインシデントと言える出来事が多数発生しており、いよいよ AI セキュリティという課題領域の重要性が実体化してきた状況です。事例の数は増加傾向にあり、ヒューマノイドに組み込まれた LLM が間接プロンプトインジェクションによりジェイルブレイクされ人にモデルガンを向け発砲したという事例や、直近で発生した銃乱射事件について X.com 上の Grok が出鱈目な応答を繰り返し混乱を招いた事例など、AI セキュリティとの関連がありつつも、今号では掲載を見送った事象も複数あります。四半期に一度や 2 ヶ月に一度というペースでは状況に追随できなくなってきておりますので、現在の形態での情報のご提供を改め、注意喚起につながる情報をなるべく迅速にお届けできる仕組みを検討中です。