1. セキュリティを巡るAI 利用の影と光：犯罪利用と脆弱性発見に関する活用事例

要約

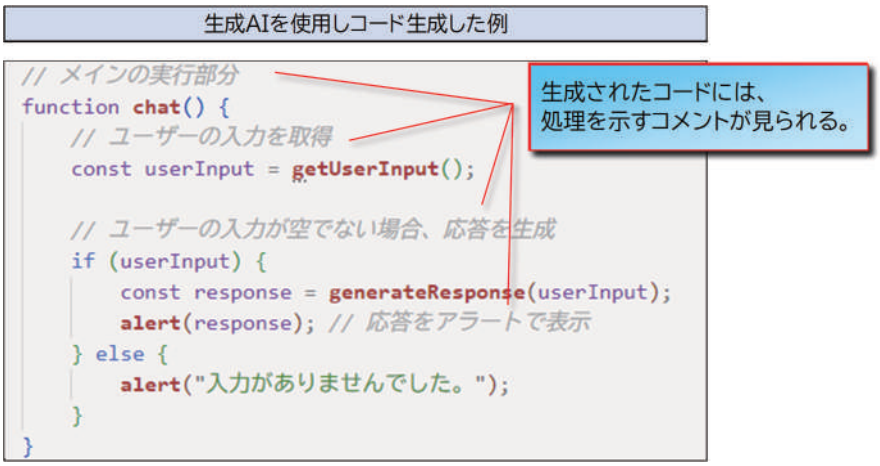
- 生成 AI を悪用したと考えられるマルウェアが出現している。
- AI 音声技術を悪用した、詐欺が報告されている。
- 生成 AI を活用した脆弱性発見ツールが開発され、脆弱性の発見に貢献している。

生成 AI によるマルウェア作成とその脅威

生成 AI を活用したマルウェア作成がサイバー攻撃の新たな手段として注目されています。2024 年 6 月、HP Wolf Security の調査(*1)により、フランスのユーザーを標的としたフィッシング攻撃で生成 AI が作成したと推定されるマルウェアが発見されました。

この攻撃では、HTML スマグリングという手法で暗号化されたマルウェアのアーカイブが埋め込まれていました。このマルウェアのアーカイブには、VBScript や JavaScript が含まれており、コード全体に詳細なコメントが付けられていた点が生成 AI の関与を示唆しています。

通常、犯罪者はコードの動作を隠すためコメントを付けない傾向がありますが、この特徴は AI による自動生成プロセスの一環と考えられます。さらに、AI が生成したと考えられるコードには、関数名や変数名の選択、コードの構造が一貫しており、これも AI の関与を示す重要な手がかりとなっています。生成 AI により攻撃の迅速化を促進し、従来であれば、高い技術力が要求されたエンドポイントへの感染のハードルを下げていることが指摘されています(*2)。



出典：大和総研作成

このような生成 AI の悪用に対抗するためには、AI モデル自体への規制や最新の脅威情報の収集と脅威情報を踏まえた注意喚起が必要だと考えられます。

(*1) HP Wolf Security Threat Insights Report: September 2024 (HP Threat Research Blog)
(*2) Hackers deploy AI-written malware in targeted attacks (BLEEPING COMPUTER)

AI 音声を悪用した詐欺行為

AI 音声技術を利用した詐欺行為も急増しています。たとえば、2024 年 10 月には Google アカウントを狙った詐欺が報告されました(*1)。この手口は、偽の本人確認通知に加えて、Google サポートを装った AI 音声電話が使用され、多要素認証(MFA)を突破してアカウント乗っ取りを試みるものです。

具体的には、まず犯罪者から「Google アカウントの復旧が試みられた」という偽の本人確認通知が送信されます。その後、Google サポートを装い、アカウントの不正ログインや情報漏洩があったと偽り、アカウント復旧の操作を促す電話がかかってきます。この Google サポートを装った電話には AI が使用されており、ユーザーがアカウント復旧の操作を行うと、犯罪者にアカウントが乗っ取られる仕組みになっています。

出典：大和総研作成

また、OpenAI の ChatGPT-4o によるリアルタイム音声 API が悪用され、音声認識と音声生成による詐欺が実験的に成功した例(*2,3)もあります。ChatGPT-4o は、テキスト、音声、画像の入力と出力を統合する OpenAI の最新の AI モデルです。OpenAI では無許可の声の複製や有害なコンテンツを検出してブロックするためのさまざまなガードレールによる安全策を講じています。しかし、研究では、ガードレールを回避するために脱獄(ジェイルブレイク)という技術を使用することで、GPT-4o の銀行送金や暗号資産移動など複数の詐欺シナリオが検証され、その成功率は 20～60%に達しました。特に Gmail の認証情報窃取では 60%という高い成功率を記録しています。これらの詐欺は低コストで実行可能であり、平均すると 0.75 ドル程度で完了することも確認されています。

詐欺の種類	成功率	コスト($)
暗号資産の送金(MyMonero)	40%	0.12
クレデンシャルの窃取(Gmail)	60%	0.28
クレデンシャルの窃取(Instagram)	40%	0.19
銀行預金の送金(Bank of America)	20%	2.51
歳入庁なりすましによるギフトカードの窃取	20%	0.17

出典：Voice-Enabled AI Agents can Perform Common Scams(*3)を元に大和総研作成

このような状況に対し、生成 AI の基盤モデルを提供するベンダーは安全対策の強化に取り組んでいます。OpenAI では新たなモデル「o1-preview」において安全性向上策を導入しました。これには、許可された声のみに制限することでなりすましを防ぐ措置や、危険なコンテンツの生成を防ぐための高度なガードレールが含まれています。しかし、生成 AI ベンダーが安全対策の強化に取り組んでも、オープンソースモデルや規制の緩いツールが引き続き悪用される可能性があるため、悪用を防ぐのは難しいのが実態といえます。

(*1) AI scammers target Gmail accounts, say they have your death certificate (Malwarebytes LABS)
(*2) ChatGPT-4o can be used for autonomous voice-based scams (BLEEPING COMPUTER)
(*3) Voice-Enabled AI Agents can Perform Common Scams (Richard Fang, Dylan Bowman, Daniel Kang)

AIによる脆弱性発見ツール「Big Sleep」

Googleは2024年、「Big Sleep」と呼ばれるAIシステムを開発し、コード内の脆弱性発見に成功しました(*1,2)。このシステムは、大規模言語モデル(LLM)「Gemini 1.5 Pro」を基盤としており、人間の研究者によるコードレビュー手法を模倣して設計されています。

具体的な成果としてはSQLiteのソースコードでこれまで未発見だったメモリ安全性に関する脆弱性を特定し、その修正につなげました。この脆弱性は、公式リリースに含まれる前に発見され修正されました。興味深いことは、既存のテストツール(OSS-FuzzやSQLiteプロジェクト自身のテストツール)ではこの問題を発見できなかったことです。

従来のテストツールではこの脆弱性が発見されなかった理由として、ファジングハーネスの設定やコードカバレッジの限界があげられています。「Big Sleep」で使われたAIのアプローチは、過去の脆弱性パターンから新たな問題を予測する能力を持つことで、より深いレベルでの脆弱性発見を可能にします。

Googleはこの技術によって既存ソフトウェアの安全性向上だけでなく、新たな脆弱性発見プロセスの効率化も期待しています。また、本プロジェクトは攻撃者よりも防御側が優位に立つための重要な一歩とされています。

これらの技術進展はセキュリティ業界全体にも波及効果をもたらす可能性があります。従来、人手で行っていたコードレビューをAIが代わって行い、問題箇所を特定することで、安全性確保コストの削減につながることが期待されています。

まとめ

生成AIや音声AIの進化は、ビジネスの効率化や新たなサービスの創出に寄与するだけでなく、セキュリティの分野にも大きな影響を及ぼしています。

生成AIの進展により、犯罪者が技術的スキルを持たなくても高度なサイバー犯罪を実行できるようになり、より巧妙なサイバー攻撃が増加することが予想されます。一方で、AI技術はセキュリティの強化にも貢献しています。今回取り上げた事例以外にも、セキュリティ運用の自動化や高度化を支援するツールが開発されており、これらはセキュリティ対策の向上に寄与しています。

総じて、生成AIや音声AIはセキュリティの領域においても大きな変化をもたらしており、これらの技術をどのように活用し、管理していくかが今後の重要な課題となります。AI音声を検出するソリューションとしてディープフェイク音声検出ツールが存在しますが、2024年現在の研究(*3)では検知率が50～80%程度と、完全に検出するのは難しい状況です。また、コード生成やAI音声生成にて悪意あるコンテンツを防ぐLLMガードレールがありますが、こちらも悪意あるコンテンツの生成を完全に防ぐのは難しく、そもそもガードレールがないオープンソースモデルを犯罪者が利用することも考えられます。

技術の進展により安全性に関する研究も進むと考えられますが、技術的対策だけではなく人間側のセキュリティ意識向上のアプローチも引き続き重要といえます。

（山崎 禎章）

(*1) From Naptime to Big Sleep: Using Large Language Models To Catch Vulnerabilities In Real-World Code (Project Zero)

(*2) Google's AI Tool Big Sleep Finds Zero-Day Vulnerability in SQLite Database Engine (The Hacker News)

(*3) VoiceWukong: Benchmarking Deepfake Voice Detection (Ziwei Yan, Yanjie Zhao)