（報道発表資料）

2026.11.16

NTT 株式会社

ソブリン型でフロンティア AI に迫るサイバーセキュリティ AI 技術を開発

～外部にソースコードを出すことなく、高い脆弱性発見能力を実現～

発表のポイント：

- ◆ 機密情報である検査対象のソースコードを外部 AI サービスに送信せず、自ら管理する環境において脆弱性を解析可能とするサイバーセキュリティ AI 技術の開発を進めています。
- ◆ 今回、複数のオープンウェイトモデルに、脆弱性研究の知見を取り入れた解析技術とワークフロー制御の独自技術を組み合わせることによって解析能力を向上させ、ベンチマーク評価により、従来技術を上回る発見性能を実現しました。
- ◆ 今後は、脆弱性の発見から組織固有のリスクに応じた判断、適切な対処までを一貫して支援する AI へと発展させ、組織におけるセキュリティリスク低減をめざします

NTT 株式会社（本社：東京都千代田区、代表取締役社長：島田 明、以下「NTT」）は、検査対象のソースコードを外部 AI サービスに送信せず、高度な脆弱性解析を行う AI 技術を開発しました。複数のオープンウェイトモデルと、脆弱性研究の知見を取り入れた解析技術とワークフロー制御の独自技術を組み合わせることにより、脆弱性解析のベンチマーク評価において、従来手法の 48%から 63%へと成功率を 15 ポイント向上させました。これにより、高度な解析能力と、組織が AI を自ら選択・運用できるソブリン性の両立をめざします。

今後は、本成果を基盤に、脆弱性の発見から組織固有のリスクに応じた判断、対処までを一貫して安全かつ迅速に支援するサイバーセキュリティ AI 技術の実現をめざします。

本研究成果は、NTT R&D FORUM 2026（※1）に展示予定です。

1. 背景

AI の急速な進化により、ソフトウェアを解析し、脆弱性を発見する能力が大きく向上しています。こうした能力がサイバー攻撃にも悪用される懸念が高まる中、守る側にも AI を活用した迅速な対応が求められています。

一方、高性能な外部 AI サービスの利用においては、解析対象のソースコードなどの機密情報の外部送信や利用コストの増大、特定サービスへの依存といった懸念があります。

また、脆弱性が発見されても、それだけでセキュリティリスクが低減するわけではありません。多数の脆弱性から自組織のシステムに影響するものを見極め、優先順位を判断し、安全に対処する必要があります。発見能力の向上に伴い、こうした判断・対処の迅速化も重要になります。

NTT は、AI のモデルと独自の解析技術と解析ワークフローを組み合わせながら、組織が AI の選択・運用に関する主導権を持つ「ソブリン性」を確保しながら、脆弱性の発見から判断・対処までを一貫して支援するサイバーセキュリティ AI 技術の実現をめざして研究開発を進めています。（図 1）

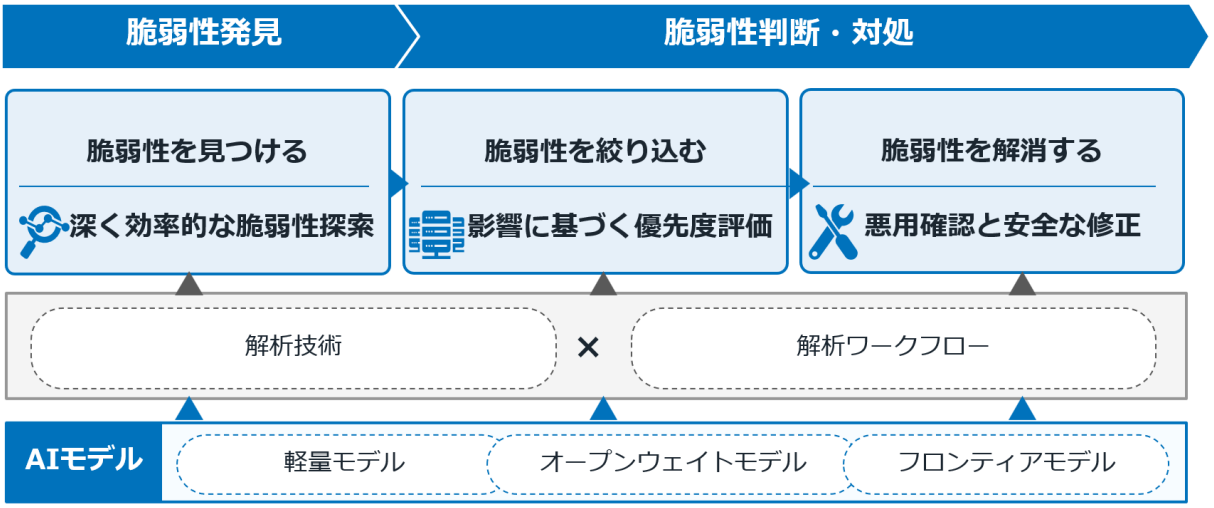


図1. サイバーセキュリティ AI 技術の全体像

2. 技術のポイント

NTT は、独自の解析技術と AI エージェントの解析ワークフローを改良することによって、ソースコードを外部 AI サービスに送信せずに利用できる技術の解析能力を向上させました。

一般的な AI を活用した脆弱性発見では、AI モデルがプログラムを読み解き、脆弱性の可能性について仮説を立てて解析し、ねらいを定めて検証します。専門家による調査に近い一連の作業を、AI は自律的に繰り返し、並列して大量に実行できることが特徴です。

こうした AI の脆弱性解析能力は、モデル自体の性能だけでなく、解析の進め方を定める「ワークフロー」や検証に使用する「ツール」によっても大きく左右されます。

NTT は、これまでの脆弱性研究で蓄積した知見を活用し、脆弱性が生じる可能性について AI に着眼点を与える手引きを解析ワークフローに組み込みました。これにより、AI の能力を効果的に引き出し、脆弱性解析の成功率を向上させました。(図 2)

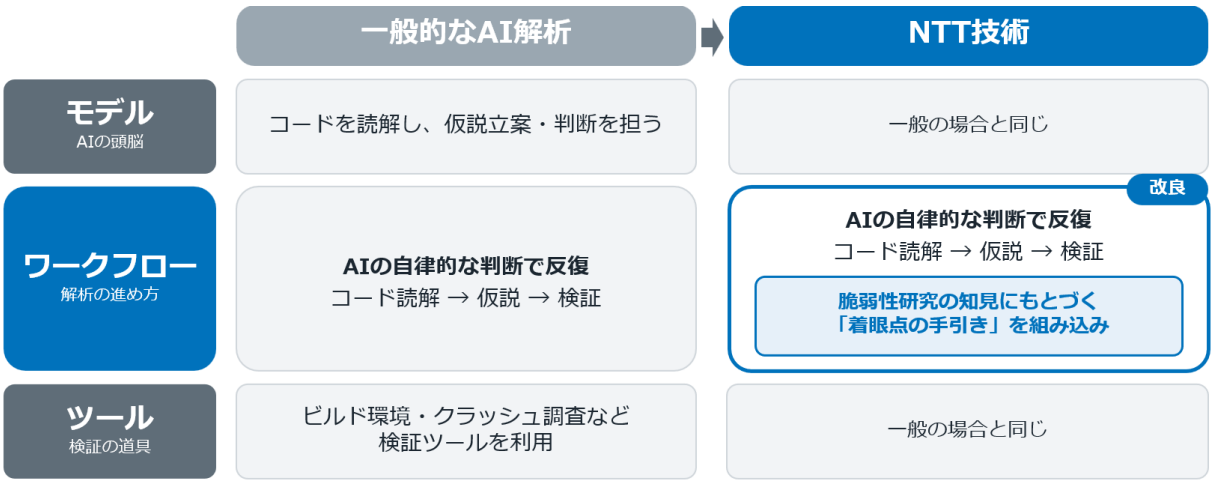


図 2. 一般的な技術と NTT 技術の比較

これまで NTT では、検査対象への入力を繰り返し変え、異常な動作から脆弱性候補を発見する

解析技術についての研究開発を進めてきました。今後は、NTT が培ってきた解析技術との連携も進め、発見能力のさらなる向上に取り組みます。更に、解析能力と計算資源のバランスを考慮し、組織が管理する環境でも利用しやすい技術へと発展させていきます。

3. 実験の概要

今回の評価では、AI エージェントの脆弱性解析能力を評価するベンチマーク「CyberGym」（※2）（※3）の約 1,500 件の課題から、メモリ破壊に関する 150 件を選定し、既存の解析ワークフローと今回の改良手法を比較しました。

その結果、既存手法では 48%だった解析成功率が、改良手法では 63%となり、15 ポイント向上しました。AI モデル自体の性能だけでなく、解析の進め方を工夫することで、脆弱性解析能力を高められることを確認しました。

また、規模の異なるオープンウェイトモデルを用いた評価を実施することによって、中規模モデルの組合せによる性能向上の可能性を確認しました。

4. 今後の計画

NTT は、今回の成果を基盤に、組織固有のリスクに応じた判断から対処までを一貫して安全かつ迅速に支援するサイバーセキュリティ AI 技術の開発を進めます。

各組織においては、自ら発見した脆弱性に加え、日々外部で公表される膨大な脆弱性への対応も求められます。しかし、そのすべてが自組織システムに影響するわけではなく、対処の必要性や優先順位も異なります。そこで、脆弱性情報とシステム構成などを照合し、影響を受ける箇所や優先して対処すべきリスクを絞り込み、システムの運用状況や業務への影響を考慮した対処を支援することによって、脆弱性判断・対処を強化する AI 技術の研究開発を進めます。

これらの技術を連携させることによって、従来は専門家が個別に実施してきた脆弱性の発見・判断・対処を一貫して支援し、実際のリスク低減までに要する時間と負担の削減につなげていきます。

5. 展開

今後は、高度な AI を、自ら管理する環境で活用できる技術の展開を通じて、迅速な脆弱性対応と安全なシステム運用に貢献していきます。

【用語解説】

※1 NTT R&D FORUM 2026 https://www.rd.ntt/forum/2026/

Mehr Licht!
More Photonics for Intelligence

※2 GitHub - sunblaze-ucb/cybergym: CyberGym is a large-scale, high-quality cybersecurity evaluation framework designed to rigorously assess the capabilities of AI agents on real-world vulnerability analysis tasks. · GitHub

※3 CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale

■本件に関する報道機関からのお問い合わせ先
NTT 株式会社
サービスイノベーション総合研究所
広報担当
問い合わせフォームへ