Nomura Research Institute Group

NEWS RELEASE

2024 年 6 月 19 日

NRI セキュアテクノロジーズ株式会社

NRI セキュア、生成 AI を組み込んだシステム向けの セキュリティ監視サービス「AI Blue Team」を提供開始

NRI セキュアテクノロジーズ株式会社（本社：東京都千代田区、代表取締役社長：建脇 俊一、以下「NRI セキュア」）は、「AI セキュリティ統制支援サービス」[1]のラインナップの一つとして、生成 AI を活用したシステムを対象にセキュリティ監視を行う「AI Blue Team（以下「本サービス」）」の提供を、本日開始します。

2023 年 12 月にリリースしたセキュリティ診断サービス「AI Red Team」[2]で、システム固有の脆弱性を洗い出したうえで本サービスを利用することで、大規模言語モデル（LLM）[3]を活用したシステムのセキュリティ対策を包括的かつ継続的に実施することができます。

■ 生成 AI 導入におけるリスク

生成 AI、特に LLM を情報システムに組み込んで業務効率化や新規サービスに活用する動きが急速に広がる中で、それらについてのセキュリティ対策がますます重要になっています。LLM は、「プロンプトインジェクション」[4]「プロンプトリーキング」[5]等の攻撃の標的となりうる脆弱性のほか、「ハルシネーション」[6]「不適切なコンテンツの生成」「バイアスリスク」[7]等の生成内容の信ぴょう性の観点や、予期せぬ「機微情報の漏洩」といったインシデント等、多岐にわたるセキュリティリスクをはらんでいます。

■ 本サービスの概要と特長

NRI セキュアは、生成 AI（LLM）を取り巻くこれらのリスクを的確に検知することに加え、新種の攻撃手法やシステムの運用開始後に検出された脆弱性に対処するため、関連情報を「インテリジェンス（セキュリティ脅威について収集・分析した情報等の総称）」として継続的に蓄積し、監視業務に適用していくことを重視しています。このたび、LLM に関連するリスクから導入企業のシステムを保護する仕組みを独自に開発し、アプリケーションとして本サービスに組み込みました（特許出願中）。

本サービスの目的は、企業が LLM の活用による業務効率化やビジネスの変革に注力できるよう、継続的な監視を通じて LLM に関連するリスク管理を支援することです。

なお、本サービスの導入にあたって、はじめに AI Red Team によるセキュリティ診断を実施します。AI Red Team の診断結果を元に、システム固有の問題に対するインテリジェンスを本サービスのアプリケーションに適用することで、一般的な AI 防御ソリューションでは対処することが難しいリスクにも対策を施すことができます。

本サービスの主な特長は、以下の 2 点です。

1. 生成 AI 活用システムを継続的にモニタリングし、広範囲かつ最新の AI リスクを回避

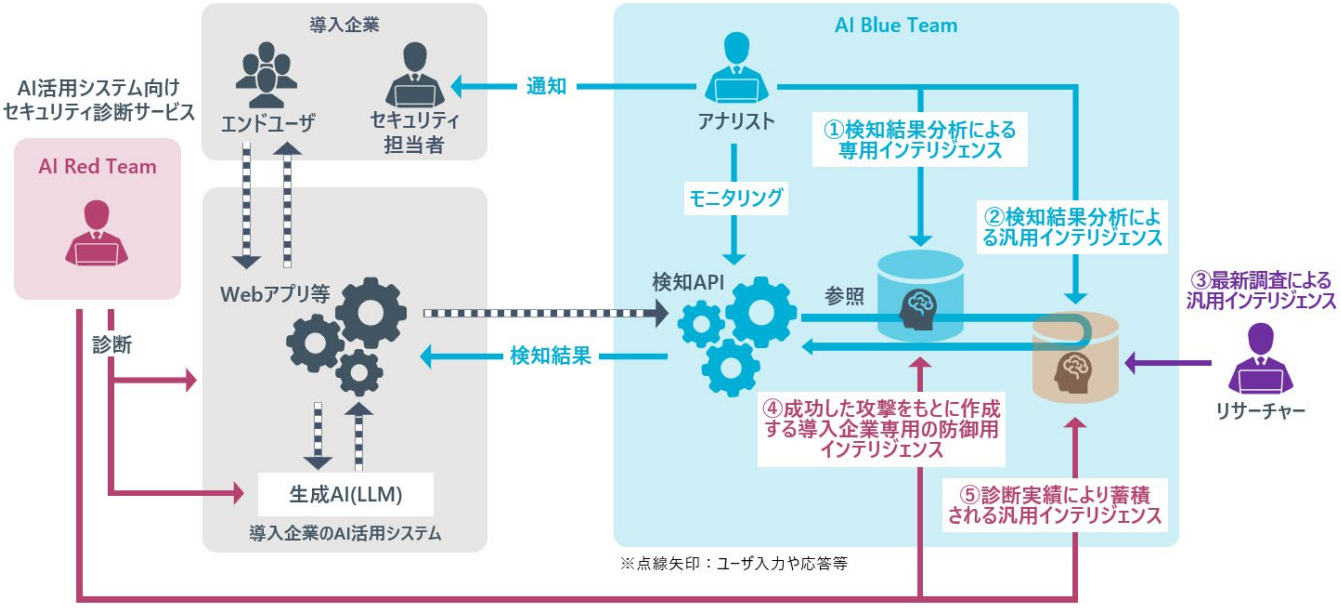
監視対象となるシステムと生成 AI 間で行われる入出力の情報を、本サービスで提供する「検知 API」[8] に連携し、有害な入出力を検知した場合、導入企業の担当者に通知します。前述した LLM が抱えるセキュリティリスクのモニタリングと対応に留まらず、NRI セキュアのアナリストが検知結果をもとに攻撃傾向を分析し、新種の攻撃手法等にも対応できるようにインテリジェンスを蓄積し、継続的に最新化します。アナリストがモニタリングする監視用画面は、導入企業の担当者もアクセスできるため、検知状況を直接確認することも可能です。

2. AI Red Team で検出したシステム固有の脆弱性を防御し保護レベルを強化

生成 AI を活用したシステムの開発現場では、どのようなケースで AI を活用しているか等の AI への依存方式や権限委譲レベルに応じて、システム固有の脆弱性が作り込まれてしまうことがあります。このような脆弱性は汎用的な防御ソリューションでは対処できず、個別の対策が求められます。

本サービスでは、AI Red Team のセキュリティ診断で検出されたシステム固有の脆弱性にも対応できるよう、汎用的なインテリジェンスとは別に導入企業専用のインテリジェンスを蓄積します。これにより、固有の脆弱性を攻撃から防御しつつ、システム全体の保護レベルを一層強化することが期待できます。

図：AI Blue Team と AI Red Team を併用した生成 AI 活用システムのセキュリティ強化策



本サービスの詳細については、次の Web サイトをご参照ください。

https://www.nri-secure.co.jp/service/assessment/ai-blue-team

NRI セキュアは今後も、企業・組織の情報セキュリティ対策を支援するさまざまな製品・サービスを提供し、安全・安心な情報システム環境と社会の実現に貢献していきます。

---

[1] AI セキュリティ統制支援サービス：詳細については「ご参考」を参照ください。

[2] AI Red Team：新たに生成 AI を利用するシステムやサービスを対象にした、セキュリティ診断サービスです。詳細は次の Web サイトをご参照ください。https://www.nri-secure.co.jp/service/assessment/ai-red-team

[3] 大規模言語モデル（LLM）：LLM は、Large Language Model の略で、大量のテキストデータを利用してトレーニングされた自然言語処理モデルのことです。

[4] プロンプトインジェクション：主に、攻撃者が入力プロンプトを操作して、モデルから予期しない、または不適切な情報を取得する試みを指します。

[5] プロンプトリーキング：攻撃者が入力プロンプトを操作して、もともと LLM に設定されていた指令や機密情報を盗み出そうとする試みを指します。

[6] ハルシネーション：AI が事実に基づかない情報を生成する現象を指します。

[7] バイアスリスク：トレーニングデータの偏りやアルゴリズム設計により、偏った判断や予測を引き起こす現象を指します。

[8] API：Application Programming Interface の略称で、プログラムの特定の機能をその他のプログラムでも利用できるようにする技術です。

【ニュースリリースに関するお問い合わせ先】

NRI セキュアテクノロジーズ株式会社　広報担当

TEL：03-6706-0622　E-mail：info@nri-secure.co.jp

【ご参考】

NRI セキュアが提供する AI セキュリティ統制支援サービスについては、次の Web サイトおよび一覧表をご参照ください。https://www.nri-secure.co.jp/service/ai-security

「AI セキュリティ統制支援サービス」のメニュー一覧

サービスメニュー	サービス概要	支援対象フェーズ	提供状況
AIリスクガバナンス構築支援サービス	AI利活用推進におけるガバナンス対応全般（AI利用ポリシーやガイドライン整備、データガバナンス策定等）に関する、コンサルタントによる人的な態勢整備を支援するサービス	企画構想～開発～導入・運用	提供中
AI品質・適合性検証サービス	欧州AI規制やISO標準規格等、リスクベース規格への適合性検証に関わる技術評価を行うことでAIシステムの品質を評価するサービス	開発～導入・運用	提供中
AIセキュリティ診断サービス (AI Red Team)	攻撃者の視点で、AIシステムを攻撃し、潜在リスクを洗い出して改善に寄与するサービス	導入～運用	提供中
AIセキュリティ監視サービス (AI Blue Team)	AIシステムを独自のメソッドでモニタリングし攻撃を検出・防御することで安定運用に寄与するサービス	運用	今回提供開始
セキュアAI基盤構築支援サービス	企業がAIシステムを安全に利用するにあたって、各種セキュリティ対策を施した専用環境を提供、構築を支援するサービス	開発	2024年度中開始予定
データセキュリティサービス	AIシステム内のデータの管理態勢・データ流通を可視化するソリューションの提供によりデータ管理に関するビジネスリスクを軽減するサービス	開発～運用	2024年度中開始予定