特集／AIとサイバーセキュリティを巡る動向と対応

〈4〉AIシステムのサイバーセキュリティ上の問題について

～システムを支えるインフラの脆弱性悪用の攻撃傾向から～

一般社団法人 JPCERT コーディネーションセンター
脅威アナリスト　**佐々木　勇人**

はじめに

AIをめぐるサイバーセキュリティ上の問題については、すでに多くの調査・研究が行われている。「AIを悪用した攻撃」に関する注目度は高いものの、すでに多数発生している「AIシステム／サービス自身が稼働するシステム自体におけるインシデント」については、あまり注目されていない。こうした、AIシステム／サービスが稼働するソフトウェア／プラットフォームを取り巻くセキュリティ上の問題点やインシデント事例について解説[1]する。

AIをめぐる「セキュリティ上の問題」については、大別すると

① AIの「使い方」の問題：AIを用いてサイバー攻撃などの不法行為を行うこと

② AIモデル自体（アルゴリズム）のセキュリティ上の問題：例）AIが意図しない回答をしてしまい、本来外部に漏えいしてはならない情報が開示されてしまう等

③ AIモデル自体（ソフトウェア）のセキュリティ上の問題：オンラインまたはオンプレミスで提供される、ソフトウェア／サービスとしてのAIモデルにソフトウェア上の脆弱性が存在すること

④そもそもセキュリティ機能が実装されていない問題：多くのソフトウェアでは実装されているセキュリティ対策が新興のAIソフト／サービスでは実装されていないこと

⑤上記以外のセキュリティ上問題：AIモデルを実装したシステム／サービス全体やインフラのセキュリティ上の問題、API、プラグイン等のソフトウェア周辺部・連携するアプリケーション側の脆弱性等

と整理することができる。「AIのセキュリティ」というと「AIモデルのセキュリティ」と捉えられることが多く、上記①、②について解説したコンテンツが散見される。もちろん、こうしたトピックも重要であるが、本稿では、実際にセキュリティ上のインシデントやトラブル事例がすでに出始めている、③～⑤（特に⑤）について解説する。

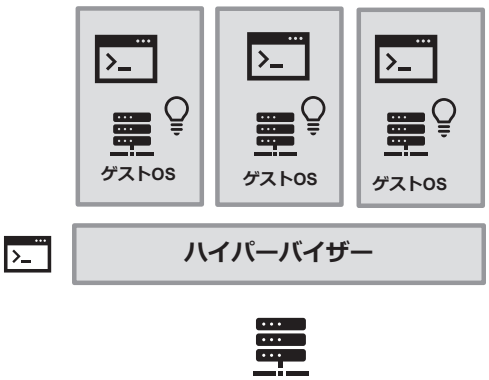
[1] ※本稿は、筆者が理事を務めるサイバーセキュリティ法制学会の研究会で発表した内容を編集・追加して執筆したものである。

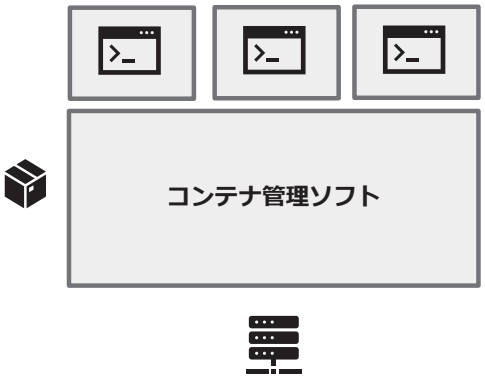
AI システムを支えるインフラ：コンテナの活用

AI サービスに限らず、多くのオンラインサービスはクラウド上のインフラで稼働している。特に、多くのユーザーにサービスを提供するものや、大規模なデータ処理を伴うサービスについては、その処理のための膨大なリソースを賄うためクラウドサービスが多用されている。特に最近では「コンテナ」と呼ばれる、サービスに必要なさまざまなソフトウェアをパッケージで管理・提供するシステムが急増している。

クラウドに限らず、多くのシステムは「仮想環境」で提供されているが、従前の仮想環境は下記図左側のとおり、仮想マシンと呼ばれる、一つ一つのホストを立て、その中でアプリケーションを動かす仕組みが採られてきた。コンテナでは下記図右のとおり、OS を一つ一つ立てるのではなく、ミドルウェアからアプリケーションまでを一つのコンテナとして稼働しており、これによって従前の仮想環境よりも素早い起動やコンピューターリソースの効率化・コスト削減を図っているのである。また、スケーラビリティ（必要に応じてコンテナを追加して拡張することができる）の点でも従前の環境よりも優れているとされており、多くのコンピューターリソースが必要であり、かつ、多数のユーザーを相手とする AI サービスにはよく用いられている。

余談であるが、最近ではサイバー攻撃の攻撃者もインフラとしてコンテナを用いるケースが出てきており[2]、やはりこちらも効率的なインフラ運用やスケーラビリティの観点からコンテナを採用していると思われる。





AI システムはそれ単独で一つのサービスを完結するというよりも、組織内外のさまざまな IT システムと連携して使われることが多く、こうしたシステムとの間は基本的に API で連携されることが多い。こうした API の機能は AI システムに限らず、脆弱性や脆弱性以外の管理不備を突いた不正なアクセス先として狙われることが多く、AI システム関連のソフトウェアでもすでに問題が出始めている。

例えば、2023 年 11 月に機械学習用のオープンソースのフレームワークである「Ray」のジョブ API における認証欠如の脆弱性（CVE-2023-48022）が公開された。IoC 検索エンジンを運営する Censys 社は当該脆弱性が残留するホストに関する調査レポートを公表[3]し、315 台のホストが公開状態にあると指摘した。他方で Ray の開発側は当初、脆弱性とは認めず、脆弱性が指摘されているダッシュボードについては、そもそも外部からのアクセスが制限されているのが運用の前提であるとの見解を示していた。

前述のとおり、こうした API に係る脆弱性や管理不備の問題は AI システムに限らずさまざまなソフトウェアやプラットフォームで度々見つかり、悪用も多数発生しているわけであるが、先に紹介したコンテナを活用したシステムでは、特にインターネット経由でのデータのやり取りが攻撃ポイントとなってしまうケースが相次いでいる。

[2] https://blog.talosintelligence.com/new-persistent-attacks-japan/
[3] https://censys.com/advisory/cve-2023-48022

コンテナとレジストリ

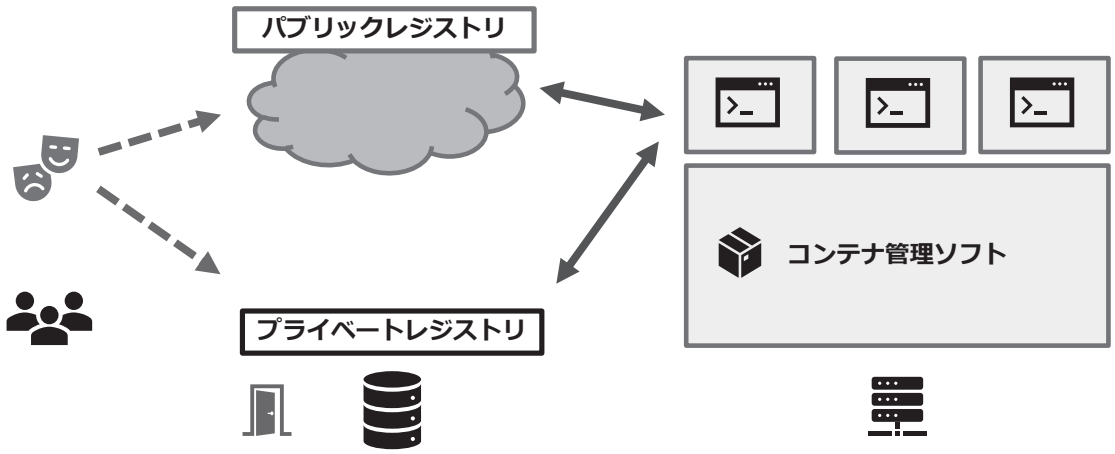
コンテナはそのスケーラビリティが特徴・メリットの一つであると解説したが、アプリのパッケージを外部からダウンロードすることで、どのようなインフラ上でもすぐに必要なソフトウェア環境が用意できる点（ポータビリティ）とスケーラビリティはセットで利点として認知されている。

こうしたパッケージは「イメージ」（コンテナイメージ）と呼ばれ、「レジストリ」と呼ばれるオンラインあるいはオンプレミスのサーバーに格納・管理され、必要な時にダウンロードされる。これは何も「単に部品を外に置いている」ということではなく、レジストリ上で開発・保守・バージョン管理などを行い、ソフトウェアのライフサイクルプロセスを維持しているのである。

レジストリには 2 種類ある。誰でも使えるコンテナイメージが保存・公開されるパブリックレジストリと、あくまでもそのコンテナイメージを使う組織だけがアクセスできるプライベートレジストリである。実際には、他組織に公開する意図がないにもかかわらず管理不備により外部レジストリを公開してしまっていたり、認証不備によりイメージの勝手なダウンロードだけでなく不正プログラムが挿入された「汚染されたイメージ」が勝手にアップロード（プッシュと呼ばれる）されたりするケース[4]も出てきている。

2024 年にトレンドマイクロ社が調査[5]したところ、（意図されないものも含めて）公開されているレジストリのほとんどでイメージを外部からプッシュできてしまう状況であった。公開されてしまっていた AI モデルのイメージの大半がオープンソースフォーマット：ONNX（Open Neural Network Exchange）によるものであり、ONNX における整合性チェックの不備を突き、検出されないような改ざんをイメージに仕込むことが可能との指摘がなされている。

意図せず公開されているイメージがダウンロードされた場合、イメージ内部にある機微な情報が外部に漏えいすることとなるが、逆に改ざんされたイメージがプッシュされるとどのような被害につながるのであろうか。



2024 年 9 月に公表された、NVIDIA Container Toolkit のコンテナエスケープの脆弱性（CVE-2024-0132）は、細工されたコンテナイメージによって、攻撃者は特定のコンテナから“抜け出し”、ホストシステムや他のアプリケーションへの侵害が可能となる脆弱性[6]であった。このソフトウェアは GPU を活用したコンテナを構築するために、多くの AI サービスで採用されており（AWS、Azure Kubenetes

[4] https://www.trendmicro.com/vinfo/us/security/news/virtualization-and-cloud/exposed-container-registries-a-potential-vector-for-supply-chain-attacks
[5] https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/silent-sabotage-weaponizing-ai-models-in-exposed-containers
[6] https://www.trendmicro.com/ja_jp/research/24/i/nvidia-ai-container-toolkit-vulnerability-fix.html