2026

AIセキュリティ
実態レポート

1,200社を超える本番クラウド環境から見えたAIセキュリティの実態

orca | RESEARCH POD

©2026 ORCA SECURITY. ALL RIGHTS RESERVED.

AIの登場により、本番環境にコードを届ける従業員は爆発的に増えました。しかし、セキュリティチームはその速度に追いついていません。人々のつくり方が変わるなら、セキュリティの働き方も変わらなければなりません。本レポートはそのギャップを測るものです。そして、ギャップを埋める出発点は、AIリスクを理解して行動するための文脈をセキュリティチームに渡すことにあります。

GIL GERON | ORCA SECURITY CEO兼共同創業者

本レポートの構成

はじめに 01

Orca Research Podについて 02

エグゼクティブサマリー 03

主要な調査結果 04

1. AI導入：実験から本番インフラへ 07

2. 攻撃を受けるAIサプライチェーン 16

3. AIエージェントとRAG：ガバナンスなき攻撃対象領域 25

4. AIスプロールとガバナンスのギャップ 30

5. AIの認証情報と安全でないアクセス 35

6. AIインフラの露出 37

7. AIの暗号化：欠落したレイヤー 42

8. 自社にとって何を意味するのか 46

9. 主要な推奨事項 48

10. まとめ 52

Orca Securityについて 53



2026年版 AIセキュリティ実態レポート

はじめに

AIは、組織がクラウドで構築し、デプロイし、運用する方法を根本から変えました。しかし、セキュリティはその変化に追いついていません。この 2年でAIは実験的なパイロットから本番インフラへと移行し、基盤モデル、自律型エージェント、 RAGパイプラインが機密データに接続されるようになりました。これらを統制するためのコントロールは存在します。ところが、その適用はデプロイに遅れ、最小権限、認証、ネットワーク分離といった基本的な衛生管理が、スピードを優先して省略され続けています。

AIワークロードはデフォルト設定のまま稼働し、APIキーは平文で置かれ、AIパッケージの重大な脆弱性は修正が存在していても未適用のまま残されています。AIで開発を行う組織の半数以上が、クラウド権限を持つエージェントフレームワークを、ガードレールなしでデプロイしています。これらは理論上のリスクではなく、構造的な現実です。

脅威の状況も変化しました。2025年から2026年にかけて、AI固有のパッケージを狙ったサプライチェーン攻撃、ゼロデイを自律的に発見・悪用するモデル、安全制御が破られた際に実害を引き起こすエージェント型システムが現れています。開発を加速させるツールが、そのまま攻撃対象領域の拡大も加速させているのです。

本レポートは、Orca Research Podによる1,200社を超える本番環境の分析に基づき、AI導入とAIセキュリティ導入のギャップを定量化し、組織が設定できていないコントロールを具体的に示し、規制が対応を強いる前にそのギャップを埋めるための実践的な指針を提示します。

Gil Geron | Orca Security CEO兼共同創業者

Orca Research Podについて

Orca Research Podは、セキュリティリスクと脆弱性を発見・分析し、Orca Securityプラットフォームの強化とクラウドセキュリティのベストプラクティスの前進に取り組むセキュリティ研究者チームです。同チームは独自の調査研究を継続的に公開しており、その成果はセキュリティ業界で広く取り上げられています。本レポートでも、より広範な調査テレメトリと分析とあわせて、その知見を随所で参照しています。

Research Podは本番環境のテレメトリ分析にとどまらず、AI固有の攻撃経路に関する独自の攻撃的セキュリティ研究も行っています。近年の公開研究には、GitHub Codespaces経由でリポジトリの完全な乗っ取りに至る受動的プロンプトインジェクションの連鎖を実証した「RoguePilot: Exploiting GitHub Copilot for a Repository Takeover」、そしてAIレイヤーを新たな侵入の起点、すなわちネットワークとアイデンティティに続く第3の水平移動の次元として提示した「AI-Induced Lateral Movement（AILM）」があります。これらの成果は、AIの攻撃対象領域がいかに速く進化しているかを示しています。

調査手法

本レポートは、Orca Securityのクラウドプラットフォームを利用する1,200社を超える本番環境から取得した、集計・匿名化済みのセキュリティテレメトリに基づいています。各指標は、データセット全体で重み付けしたうえで、該当する事象が確認された組織の割合を示しています。データは2026年第2四半期に、実際のセキュリティ体制を反映させるため本番環境のみを対象として収集しました。

調査対象は、AIクラウドサービスの設定、パッケージの脆弱性管理、シークレットの露出、エージェントおよびRAGインフラのセキュリティ、暗号化の状況、AIワークロードのアイデンティティ管理に及びます。

本レポートのデータセット

- 1,200社を超える組織のクラウドワークロードおよび設定データ
- モデル、パッケージ、エージェント、ベクターデータベース、インフラ設定を対象としたAI固有のテレメトリ
- 本レポートで参照するデータは2026年第2四半期に収集

エグゼクティブサマリー

AIは、多くのセキュリティプログラムが追随できる速度を超えて、パイロットから本番へと移行しました。本レポートは1,200社を超える組織の実環境テレメトリに基づき、AI導入のスピードとAIセキュリティの成熟度との間に広がるギャップを明らかにします。その結果として生じているのが、設定不備、未修正の脆弱性、露出した認証情報、そして統制されないまま本番稼働するエージェントです。主要な調査結果は次のとおりです。

- **AI導入は主流に、しかしセキュリティは後回し　51.5%の組織が独自アプリケーション開発にAIを導入している一方で、SageMaker環境の80%が中核となる5つの安全でないデフォルト設定をすべて有効のまま稼働させています。**
- **AIパッケージの脆弱性は広範かつ未修正　AIパッケージを利用する組織の81%が、少なくとも1件の既知のHigh脆弱性を抱えています。公開エクスプロイトが存在するアラートは、 2024年版レポートの0.2%から50.1%へ増加。修正可能なアラートの 99.9%が未適用のまま残されています。**
- **広く露出したAI認証情報が、直接的な侵害を可能にする　OpenAI利用者の28.4%が安全でない場所にAPIキーを保管しており（Anthropic SDK利用者では40%）、Azure OpenAI環境の94%が、マネージドIDではなくキーベース認証に依存し続けています。**
- **エージェントは本番に、ガードレールは未整備　AI導入企業の 56%が本番環境でエージェントフレームワークを稼働させ、　8%の組織にまたがる786のBedrockエージェントが確認されました。一方で、Bedrock利用者の 57%は安全性ガードレールを一切設定していません。**
- **暗号化は、いずれのクラウドでも実質的に不在　AIサービスを利用する組織の87～98%が、カスタマー管理暗号鍵（CMEK）を設定していません。最も高いのはSageMakerノートブックの98.4%で、学習データ、モデル、パイプラインがプロバイダー標準の暗号化に依存したままになっています。**
- **AIスプロールがガバナンスを追い越しているAIクラウド利用者の55.3%が4種類以上のAIサービスを、19.6%が7種類以上を運用しています。この広がりに対して一貫したセキュリティポリシーを維持することは、多くの組織にとってほぼ不可能です。**
- **AI規制は2026年のカレンダーに載っているEU AI Actの高リスク義務は2026年8月2日に発効し（罰金は最大3,500万ユーロ、または全世界売上高の7%）、コロラド州の改正AI法（S.B. 26-189）は2027年1月1日に施行されます。これらの設定不備は、いまや規制上の現実問題です。**

具体的な次の一手は、第 9章「主要な推奨事項」をご覧ください。

主要な調査結果

51.5%



の組織が、独自アプリケーションの開発に AIを導入しています。

AIはもはや実験ではありません。半数を超える組織が本番でAIを稼働させており、その内容も変化しています。既製モデルの利用から、エージェント、RAGパイプライン、独自のAIインフラへと軸足が移っているのです。

81%



のAIパッケージ利用組織が、既知の脆弱性を少なくとも1件抱えています（2024年の62%から増加）。

CVSSスコアの平均は6.9から8.79へ跳ね上がり、74%が深刻度Critical（9.0以上）のCVEを少なくとも1件抱えるようになりました。AIの依存関係ツリーは、2年前より深く、より深刻になっています。

50.1%



のAIパッケージ脆弱性アラートに、公開されたエクスプロイトが存在します。 2024年の0.2%から250倍の増加です。

AI脆弱性の半数が、いまや実際に悪用可能な状態にあります。それにもかかわらず、修正が提供されているアラートの99.9%が未適用のままです。「悪用は難しい」という言い分は、もはや通用しません。

80%



のSageMaker利用組織が、デフォルト設定をすべて有効のまま稼働させています。

最も広く導入されているAIクラウドサービスは、初期状態が安全ではありません。しかも、組織がそれを設定し直すことはほとんどありません。設定内容は改善しつつあるものの、大多数はいまだ5つの安全でないデフォルト設定をすべて有効のまま運用しています。

86%



のSageMaker利用組織に、名前が推測可能なバケットが少なくとも1つ存在します。

旧来の命名パターンは総当たりで推測可能で、学習データ、モデルの成果物、パイプラインの出力を露出させます。AWSがデフォルトの命名規則をランダム化したにもかかわらず、該当する組織は2024年の45%からほぼ倍増しました。

25%〜40%



のユーザーが、Anthropic、OpenAI、HuggingFaceのいずれかで、少なくとも1つのAPIキーを安全でない場所に保管しています。

Anthropicでの露出は3倍以上に増加（13%→40%）、OpenAIも上昇（20%→28.4%）、HuggingFaceは改善しました（35%→24.9%）。露出はプロバイダーを問わず広範に及び、導入が最も速く伸びている領域で悪化しています。

主要な調査結果

98%



のSageMakerノートブックインスタンスが、カスタマー管理の暗号化を利用していません。

Azure OpenAI（91.6%）とVertex AI Models（94.8%）がこれに続きます。Vertex AIは2024年から若干改善しましたが、大多数の組織は AIワークロードについて、いまだにプロバイダー管理のデフォルト暗号化に依存しています。

56%



のAI導入企業が、本番環境でエージェントフレームワークを稼働させています

最も多く使われているのはangGraphです。加えて、クラウド権限を持つ786のBedrockエージェントと、データに接続され879のナレッジベースが確認されました。いずれも024年にはなかったカテゴリであり、AIスタックが自律的かつデータ接続型のシステムへと急速に広がっていることを示しています。

57%



のBedrock利用組織が、ガードレールを設定せずにエージェントを稼働させています。

Amazon Bedrockを利用する組織の半数以上が、安全性ガードレールを一切設定していません。その結果、 AIエージェントはコンテンツフィルタリング、根拠の検証、ポリシー制御のいずれもないまま動作しています。

64%



のAI導入企業が、ベクターデータベースを導入しています。

主要な7種のベクターデータベースにまたがるRAGの普及は、取得された文書を経路とするデータポイズニングやプロンプトインジェクションに対して、新たで分断された攻撃対象領域を生み出しています。

55.3%



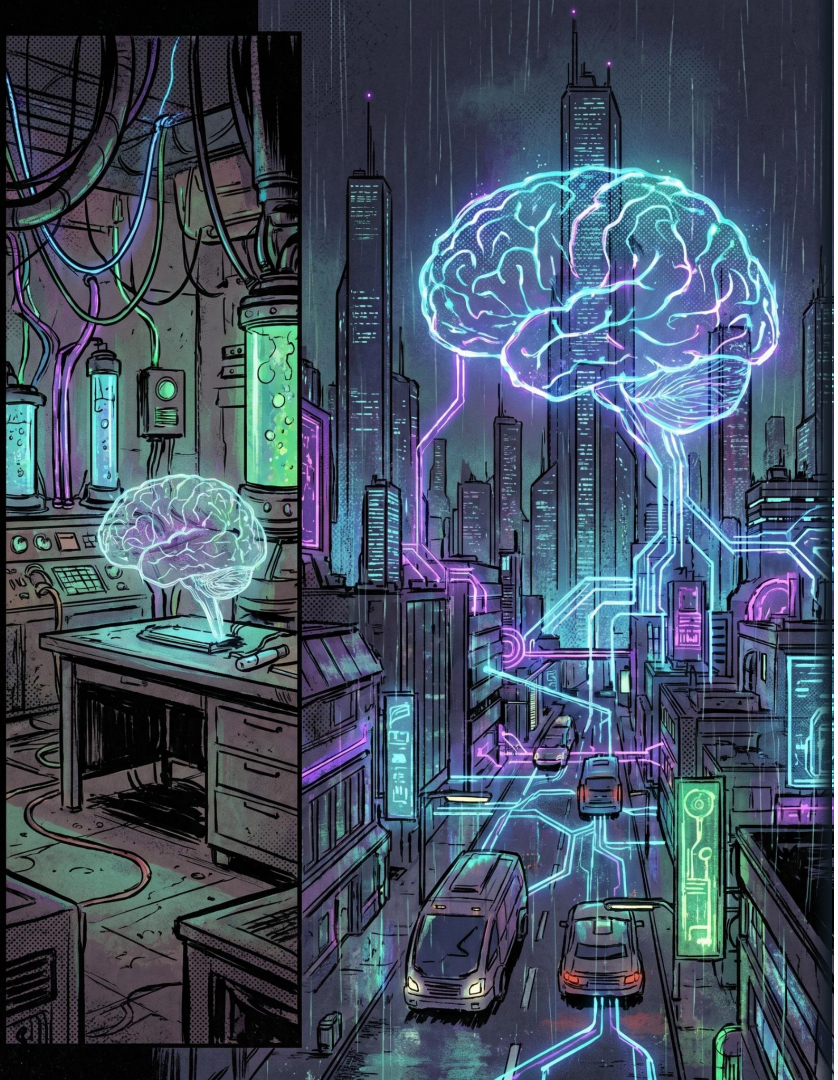
のAIクラウドサービス利用者が、 4種類以上の異なるAIサービスを運用しています。

19.6%は7種類以上を利用しています。AIワークロードが複数のサービスとクラウドに分散するなかで、一貫したセキュリティガバナンスの維持は、セキュリティチームにとって年々難しくなっています。

前年比較サマリー

指標	2024年版レポート	2026年版レポート	傾向
脆弱性のあるAIパッケージを1件以上抱える組織	62%	81%	悪化
AIパッケージのCVSS平均スコア	6.9	8.79	悪化
公開エクスプロイトが存在するAI脆弱性	0.2%	50.1%	悪化（250倍）
root権限で稼働するSageMaker	98%	75.8%	改善
IMDSv2が未設定のSageMaker	77%	47.9%	改善
デフォルトのバケット名を使用するageMaker	45%	86%	悪化
安全でない場所に保管されたOpenAIキー	20%	28.4%	悪化
安全でない場所に保管されたHuggingFaceキー	35%	24.9%	改善
安全でない場所に保管されたAnthropicキー	13%	40%	悪化（3倍）
CMEK未設定のVertex AI	98%	92.9%	小幅改善
Azure OpenAIの導入率（Azure利用組織に対する割合）	39%	50.5%	増加
SageMakerの導入率（AWS利用組織に対する割合）	29%	31%	増加
Vertex AIの導入率（GCP利用組織に対する割合）	24%	32%	増加
最も利用されているAIモデル	GPT-3.5（79%）	GPT-4o（37.6%）	世代交代
エージェントフレームワークの導入	未計測	AI導入企業の56%	新規カテゴリ
ベクターデータベースの導入	未計測	AI導入企業の64%	新規カテゴリ

01 AI導入：実験から本番インフラへ



1. AI導入：実験から本番インフラへ

2024年におけるAI導入とは、モデルをデプロイし、コパイロットをつなぎ、 APIを呼び出すことでした。企業のAIはたいてい単一の統合ポイントの背後にあり、セキュリティ上の論点はモデル自体に集中していました。

2025年、問題の形が変わりました。エージェント型 AIはデモから標準へと移行し、McKinseyの「2025 State of AI」では企業の利用率は 78%、AIエージェントを拡大中またはパイロット中の組織は62%に達しました。2024年末にAnthropicが公開したModel Context Protocolは事実上の統合レイヤーとなり、主要なモデルプロバイダーすべてが採用し、 4月までにサーバーのダウンロード数は800万件を超えました。一方で、セキュリティの状況は悪化しています。セキュリティ実務者の90%がAIツールを使用していると回答した一方、正式なコントロールを整備していた組織はわずか 32%。さらに半数以上が、AIエージェントが意図した権限を超えて動作する事象を経験していました。

2026年現在、AIはもはや「呼び出すサービス」ではありません。 51.5%の組織が独自アプリケーションの開発にAIを導入しており、AIは開発環境、セキュリティプラットフォーム、業務システム、クラウドインフラに組み込まれています。つまり AIの攻撃対象領域は、すべてのアイデンティティ、すべての統合、そして人間に代わって動作するすべてのエージェントに広がっているということです。

AI導入：実験から本番インフラへ

1.1 AI IDEの革命

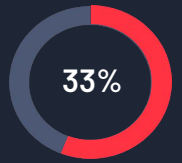
2025年から2026年にかけて、AI搭載の開発ツールはコーディング体験の標準になりました。 GitHub Copilot、Cursor、Windsurf、Claude Code、Amazon Q Developer、Gemini Code Assistが、開発者のエディタをめぐって競い合っています。Cursorはユニコーン企業となり、GitHubは自社プラットフォーム上の開発者が 1億5,000万人を超え、Copilotがワークフローに深く統合されていると報告しました。

Orcaのデータもこれを裏付けています。AI導入企業の33%がGitHub Copilotを導入し、openaiパッケージはAI導入企業の80.7%で利用されています。API経由のAIが、開発ワークフローにいかに深く組み込まれているかを示す数字です。

これらのツールは、コードベース、ターミナル、環境変数、認証情報という強力なアクセス権を持って動作します。そのアクセスが新たな攻撃対象領域を生みます。 Orca Research Podはこれを「RoguePilot」で実証しました。GitHub Codespacesに存在したこの脆弱性では、GitHub Issueに隠された受動的なプロンプトインジェクションがCopilotを密かに乗っ取り、細工されたプルリクエストをチェックアウトさせ、シンボリックリンク経由で機密の環境ファイルを読み取らせ、JSONスキーマの自動ダウンロードを通じて特権を持つ GITHUB_TOKENを外部に送 信させることで、リポジトリの完全な乗っ取りに至りました。

150 M

GitHubは、自社プラットフォーム上の開発者が1億5,000万人を超え、Copilotがワークフローに深く統合されていると報告しました。



AI導入企業の33%がGitHub Copilotを導入



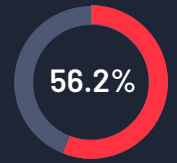
AI導入企業の80.7%。API経由のAIが開発ワークフローに深く組み込まれていることを示しています

AI導入：実験から本番インフラへ

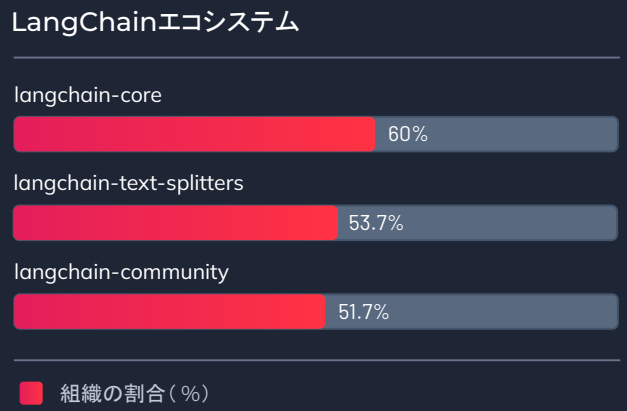
1.2 AIエージェントが本番環境へ

2026年初頭までに、AIエージェントはデモから本番へと移りました。 LangGraph、AutoGen、CrewAI、Amazon Bedrock Agents、OpenAIのAgents APIがいずれも登場、あるいは成熟しました。現在、 AI導入企業の 56.2%がエージェントフレームワークを導入しており、なかでも LangGraphが優勢です。LangChainエコシステムはエージェント型AI開発の基盤となりましたが、これは両刃の剣でもあります。成熟したツール群は導入を加速させる一方、 langchain-coreの脆弱性が 1件見つかれば、その影響範囲は本番稼働するエージェントスタックの大半に及ぶのです。

本番稼働するエージェントは、それぞれ独自の権限、メモリ、影響範囲を持つ新たな非人間アイデンティティです。そしてその多くは、セキュリティチームが棚卸しできる速度を超えて配備されています。このガバナンスギャップについては第 3章で詳しく取り上げます。



AI導入企業の56.2%が、すでにエージェントフレームワークを導入



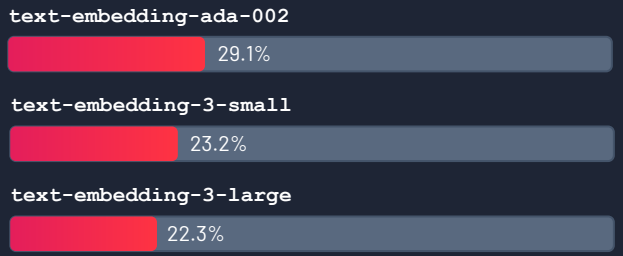
1.3 RAGと企業データの接続

組織はもはや、モデルに問い合わせるだけではありません。ソースコードリポジトリ、カスタマーサポートの会話記録、社内 Wikiといった、最も機密性の高いデータにモデルを接続しています。ベクターデータベースを基盤とする検索拡張生成（ RAG）のパイプラインによって、 LLMはクエリ実行時に社内文書、顧客データ、独自ナレッジへアクセスできるようになりました。 RAGはAIのリスクモデルを根本的に変えました。RAG以前であれば、プロンプトインジェクションは企業情報の露出につながる程度でした。RAG以後は、モデルがそれらすべてに正規の読み取り権限を持つため、プロンプトインジェクション一つで顧客記録、ソースコード、取締役会資料が外部へ持ち出され得るのです。

AI導入企業の 64%がベクターデータベースを導入していますが、その市場は 7社以上に分散しており、認証モデル、ネットワークの構成、露出に関するデフォルト設定はベンダーごとに異なります。ベンダー別の内訳と RAG固有の攻撃研究は第 3章で詳述します。企業のデータ保護ツールの多くは、これほど分断された領域を棚卸しできる前提で作られていません。エンベディング自体も機密情報です。反転（インバージョン）研究では、露出したベクターから元のテキストを部分的に復元できることが示されており、ベクターデータベースの流出は実質的に文書ストアの流出を意味します。

エンベディングモデルの利用状況は、 RAGがいかに広く使われているかを裏付けます。上位のエンベディングモデルはいずれも OpenAIのホスト型であり、つまりエンベディングの元となるテキストは、定義上すべてサードパーティの APIを経由しています。規制対象の顧客にとって、これはエンジニアリング層で暗黙のうちに下されているデータレジデンシーの判断です。

AIクラウドサービス利用者に多いエンベディングモデル



1.4 モデル勢力図の変化

クラウドサービスにおけるモデルの勢力図は分散しました。 2024年はGPT-3.5が79%の導入率で圧倒的でした。2026年時点で首位の GPT-4oはわずか 37.6%、GPT-4.1-miniが34.7%で迫っています。GPT-5ファミリーと推論モデルも、リリースから 1年足らずで実運用に入っています。かつての GPT-3.5のように市場を支配するモデルは、もう存在しません。ガバナンスへの影響は、この市場変化以上に大きいものです。1つのモデルが導入の 79%を占めていた時代は、1つのポリシーで露出の大半を覆えました。5モデルが並ぶ市場では、5組の認証情報、5種類のAPI形態、5つの安全性プロファイル、5系統の監査ログが存在します。その結果、セキュリティ制御はモデルより上位のゲートウェイ層へと移っていきます。

同時に、モデルのカスタマイズと自社ホスティングも広く行われています。従来型の機械学習や特徴量パイプライン向けに scikit-learnがAI導入企業の 82.5%で、PyTorchが63.8%で、transformersとhuggingface-hubがそれぞれ 60.8%で利用されています。いずれも、オープンウェイトモデルを自社でファインチューニングし実行するための標準的なツールチェーンです。ここから分かるのは、組織が API経由で既製モデルを利用するだけの段階を急速に脱し、独自データで基盤モデルをファインチューニングし、オープンウェイトの代替モデルを自社のクラウド環境内でホストする方向へ進んでいるということです。

これはサプライチェーンにとって重要な意味を持ちます。AI導入企業の60.8%がHugging Faceから成果物を本番環境に取り込んでいるのであれば、MLレジストリはNode Package Manager（npm）やPython Package Index（PyPI）と同じくらい重要で、同じくらい露出しているということです。2024年から2025年にかけて、Hugging Face上の悪意あるモデルの事例が複数記録されており、pickleを利用したリモートコード実行のペイロードも含まれていました。それにもかかわらず、多くの組織は本番で実行するモデルの重みよりもPythonパッケージのほうを厳密に棚卸ししています。



最も利用されているAIモデル トップ10

順位	モデル	AIモデルを導入した組織の割合
1	gpt-4o | Azure OpenAI	37.6%
2	gpt-4.1-mini | Azure OpenAI	34.7%
3	text-embedding-ada-002 | Azure OpenAI	29.1%
4	gpt-4.1 | Azure OpenAI	28.9%
5	gpt-4o-mini | Azure OpenAI	28.6%
6	text-embedding-3-small | Azure OpenAI	23.2%
7	text-embedding-3-large | Azure OpenAI	22.3%
8	gpt-5.1 | Azure OpenAI	22.3%
9	gpt-5-mini | Azure OpenAI	19.7%
10	o4-mini | Azure OpenAI	17.8%

最も利用されている AIモデル トップ10

OpenAIのモデルラインアップは2024年以降で分散しましたが、プロバイダーは分散していません。2024年のOrca「State of AI」レポートでは、当社テレメトリにおいて GPT-3.5単独でモデル導入の79%を占めていました。本レポートでは、導入率が38%を超えるモデルは1つもありません。そして、導入数の多い上位 10モデルはすべて、Azure OpenAI上で稼働するOpenAIのモデルです。

3位のtext-embedding-ada-002 は、特筆すべき結果です。 OpenAIは2024年1月に、性能が高く低コストな text-embedding-3-small とtext-embedding-3-large をリリースしました。それから 2年経っても ada-002 が29.1%の導入率を保っているのは、本レポートのサプライチェーンに関する調査結果が AIパッケージについて示したのと同じ「修正の慣性」が、モデルのバージョンにも及んでいることを示しています。本番環境のコーパスを再エンベディングする作業は運用負荷が大きく、多くの RAGパイプラインは移行を無期限に先送りしているのです。

この表からは3つの傾向が読み取れます。第一に、トップ 10のすべてが Azure OpenAI上で動くOpenAIのモデルであり、SKU単位でのモデル集中は薄れても、プロバイダー単位での集中は依然として大きいということです。第二に、トップ 10に3つのエンベディングモデルが入っていることは、 RAGパイプラインがいかに広く展開されているかを示しています。 ada-002が残り続けている点は、後継が利用可能になってからも旧世代のエンベディングが本番に居座り続けることを示唆します。第三に、GPT-5ファミリーとo4-miniは本番に入り始めた最新世代であり、その導入曲線はまだ初期段階にあります。

この表で目を引く不在は、AnthropicのClaudeモデルファミリーです。当社のテレメトリは Azure OpenAI経由の企業導入パターンを反映しています。 Claudeは主にAnthropic自身のAPIやAmazon Bedrock経由での利用が増えており、同じ導入シグナルには現れません。とはいえ、 Anthropicの企業内での存在感は大きく伸びており、第 5章の認証情報の露出に関する調査結果がそれを端的に示しています。AnthropicのAPIキーの露出は 2024年から3倍以上に増え、Anthropic SDK利用者の40%に達しました。これは当社テレメトリ上で最も高い比率です。モデル勢力図をより正確に描くのであれば、Claudeは重要かつ成長中のティア 1プロバイダーとして位置づけられます。

1.5 AI開発スタック

Orca Research Podは、組織が実際にインストールし実行している AIパッケージを追跡しています。以下のデータは、当社テレメトリで追跡しているAI／機械学習パッケージの稼働が確認された組織を母数としており、AI専用の導入と業務上付随的な導入の両方を含みます。

市場の関心が生成 AIに集まる一方で、当社テレメトリで最も多く導入されているAIパッケージは、古典的な機械学習の基盤ライブラリであるscikit-learn（82.5%）です。これは、LLMワークロードと並んで従来型MLの利用が相当な規模で存在していることを意味します。また、PyTorchが63.8%、TensorFlowが48.5%という比率は、本番 AIにPyTorchを選ぶという業界全体の流れを反映しています。 2023年以降加速してきたこの傾向が、いまでは企業テレメトリでも明確な差として現れています。

表の中位はLangChainエコシステムが占めています。langchain-core（60%）、langchain-text-splitters（53.7%）、langchain-community（51.7%）を合わせて見ると、LangChainが基本的なLLMの土台づくりからエージェントのオーケストレーションまでを担う、AIアプリケーション開発の標準的な抽象化レイヤーになっていることが分かります。2位のOpenAI Python SDK（80.7%）は、本章で先に示したモデルのデータを裏付けます。 OpenAIのエコシステムはモデル導入とSDK統合の双方で優位にあり、大多数の組織がその中で運用しているのです。

AI開発スタック

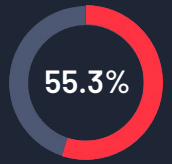
順位	パッケージ	AI導入企業に対する割合（%）
1	scikit-learn	82.5%
2	openai	80.7%
3	PyTorch（torch）	63.8%
4	transformers	60.8%
5	huggingface-hub	60.8%
6	langchain-core	60.0%
7	onnx	55.3%
8	langchain-text-splitters	53.7%
9	langchain-community	51.7%
10	tensorflow	48.5%

AI導入：実験から本番インフラへ

1.6 AIスプロール：ガバナンスの課題

AIはいまや単一のワークロードではなく、重なり合うサービス群として広がっています。 AIクラウドサービス利用者の 55.3%が4種類以上の AIサービスを、19.6%が7種類以上を運用しています。各サービスは独自のデータフロー、アイデンティティ、露出に関するデフォルト設定を伴いますが、平均的な組織にはそれらを一望できるコンソールが存在しません。

ブラウザは AI利用の主要な経路となり、同時に最も統制の及ばない領域になりました。 AI系のブラウザ拡張機能は、AI以外の拡張機能に比べて既知の脆弱性を持つ確率が 60%高く、セッション Cookieにアクセスする確率は 3倍、インストール後に権限を変更する確率は 6倍に達します。サービス、コード生成、規制上の露出にまたがるガバナンスギャップの全体像は第 4章で整理します。



AIクラウドサービス利用者の 55.3%が、4種類以上の異なる AI サービスを運用



AIクラウドサービス利用者の 19.6%が、7種類以上の異なる AI サービスを運用

02 攻撃を受けるAIサプライチェーン



2. 攻撃を受けるAIサプライチェーン

ソフトウェアサプライチェーン攻撃は、攻撃者が一度に多数の標的へ到達するための最も費用効率の高い手段になりました。2025年から2026年にかけて、この手法は AIエコシステムへ本格的に流れ込みました。ReversingLabsの調査では、Hugging Faceでダウンロード数上位 1,000のモデルのうち 23%が、いずれかの時点で侵害されていたことが判明しています。また NullifAIやModel Namespace Reuseといったキャンペーンは、攻撃者が AI固有のパッケージ、モデルハブ、エージェント関連ツールを実際に狙っていることを示しました。

Orca Research Podのテレメトリは、前節でこの露出がいかに集中しているかを示しました。この集中度では、これらのパッケージのいずれか 1つが侵害されただけで、防御側が対応する前に AI導入組織の大多数へ到達してしまいます。これが「導入とセキュリティのギャップ」の実際の姿です。 AIの依存関係は、それを棚卸し、修正し、統制するための実務の定着より速く広がっているのです。

以下の各節では、この露出の形を 5つの観点から追います。2025年から2026年の脅威を象徴した個別のインシデント、最も広く使われている AIパッケージの CVEの分布、悪用可能性の急上昇、深刻度スコアが最も高いパッケージ、そして新しい AIツールに現れ始めた脆弱性です。

攻撃を受けるAIサプライチェーン

2.1 主要なAIサプライチェーン攻撃

以下のインシデントは、AIサプライチェーンの脅威の状況を形づくったものです。通して読むと、攻撃者がAIスタックの5つのレイヤー、すなわちパッケージレジストリ、モデルハブ、開発者ツール、エージェントフレームワーク、そしてブランドへの信頼を、意図的に横断していることが分かります。これらのレイヤーはいずれも、Orcaが観測した本番環境の大多数で使われています。つまり各インシデントには、ほとんどの AI導入組織に到達し得る現実的な経路が存在するということです。事例は各ベンダーの公開情報およびCVEの報告に基づいています。

攻撃／キャンペーン		種別	影響
2026	Axios npmのバックドア1	メンテナー認証情報の窃取	北朝鮮系グループUNC1069によるOpenAIのmacOSコード署名証明書の侵害
2026	PyPI上の悪意あるlitellm2	自動実行を伴うタイポスクワッティング	偽のlitellm v1.82.8がPython起動時に自動実行され、AI／ML開発者を標的に
2026	Flowiseの重大な脆弱性3	AIワークフローツールの悪用	数千のAI環境に影響
2026	偽のClaudeサイト（PlugX RAT）4	AIブランドのなりすまし	DLLサイドローディングによるPlugXリモートアクセス型トロイの木馬の配布
2025	MCP InspectorのRCE（CVE-2025-49596）5	AIツールの脆弱性	AIエージェント開発の中核ツールにおけるリモートコード実行
2025	NullifAIキャンペーン（Hugging Face）6	プラットフォーム防御の回避	悪意あるpickle形式のモデルが、PyTorch形式の内部で標準のZIPではなく7z圧縮を用いることで、Hugging FaceのPicklescanを回避しました。既存のプラットフォーム防御が容易に突破できることを示し、pickleベースのモデル配布の見直しを迫った事例です。
2025	Model Namespace Reuse（名前空間の再利用）7	信頼モデルの悪用	攻撃者が削除済みのメンテナーアカウント名を再登録し、元の信頼された名前空間の下で人気モデルの汚染版を再公開しました。Googleは対応として、リスクのあるモデルを対象に孤立アカウントの日次スキャンを実施しています。
2025	大規模に確認されたHugging Faceの悪意あるモデル8 9 10	モデルサプライチェーンの汚染	バックドアが埋め込まれたモデル。対応としてSafetensors形式がPyTorch Foundationに参画
2026	SillyTavernのパストラバーサル11	LLMインターフェースの脆弱性	CVSS 8.8、不正なファイルアクセス
2026	PraisonAIのSSRF12	AIエージェントフレームワークの脆弱性	マルチエージェントシステムにおける未検証URLを介したSSRF

1. https://www.microsoft.com/en-us/security/blog/2026/04/01/mitigating-the-axios-npm-supply-chain-compromise/
2. https://blog.pypi.org/posts/2026-04-02-incident-report-litellm-telnyx-supply-chain-attack/
3. https://www.cisecurity.com/article/4155680/hackers-exploit-critical-flowise-flaw-affecting-thousands-of-ai-workflows.html
4. https://www.malwarebytes.com/blog/news/2026/04/fake-claude-site-installs-malware-that-gives-attackers-access-to-your-computer
5. https://www.wiz.io/vulnerability-database/cve/cve-2025-49596
6. https://www.infosecurity-magazine.com/news/malicious-ai-models-hugging-face/
7. https://unit42.paloaltonetworks.com/model-namespace-reuse/
8. https://www.ox.security/the-ox-report/ox-report-2025-retrospective
9. https://blog.gitguardian.com/silent-leak-how-malicious-hugging-face-ml-models-with-silent-backdoors/
10. https://thecyberexpress.com/hijacked-telescope-bins-python-3-boosting/
11. https://cybersecuritynews.com/sillytavern-vulnerability-cve-2026-24514/
12. https://www.linkedin.com/posts/yongchanglee_cve-2026-24486-pandas-ai-path-traversal-activity-7423...

CVEを1件以上含むAIパッケージトップ10

順位	パッケージ	AI導入企業に対する割合（%）	CVSS平均	CVSS最大
1	Pillow	92.1%	8.87	10
2	numpy	70.6%	6.69	9.8
3	scikit-learn	58.5%	7.56	9.8
4	langchain	56.4%	8.25	10
5	transformers	42.7%	8.66	9.6
6	PyTorch	42.5%	9.56	9.8
7	ray	37.8%	9.26	10.0
8	keras	37.0%	9.61	9.8
9	tensorflow	36.0%	9.76	9.9
10	onnx	31.8%	8.83	9.1

攻撃を受けるAIサプライチェーン

2.2 AIパッケージの脆弱性

AIパッケージを利用する組織の 81.2%が既知の脆弱性を少なくとも 1件抱えており、CVSSの平均は8.79です。2024年の62%、6.9から大きく悪化しました。74.1%は深刻度Critical（CVSS 9.0以上）のCVEを1件以上抱えています。実務的に言えば、本番で AIを動かしているほぼすべての組織に、モデル、学習データ、エージェント実行を扱うワークロードへ至る深刻度の高い経路が少なくとも 1本存在し、その経路は1年前より深刻になっているということです。

これらは周辺的な依存関係ではありません。Pillowはほぼすべての画像パイプラインを担い、numpyとscikit-learnはMLの中核的な数値計算を支え、PyTorchとTensorFlowはモデルそのものを実行し、LangChainは本番稼働するエージェントの多くが構築の土台とするフレームワークです。このいずれかに脆弱性があれば、それはAIワークロードのクリティカルパスの内側に入り込みます。

CVEを1件以上含むAIパッケージ トップ10

この表の要点は、古い脆弱性が残り続けているという点です。 AIパッケージは、直近 12か月に公表された CVEと並行して、5年分の未修正CVEを引き継いでいます。そして、その両方が同じ本番環境に届いているのです。

AIパッケージで最も多く見られるCVEトップ10

numpyのCVE-2021-34141は5年前に公表されたものですが、いまなお AI導入企業の 46.9%に影響しています。Log4Shellが公表から4年近く経ってもCISAの「Known Exploited Vulnerabilities」カタログに残っているのと同じ理由です。脆弱なライブラリがいったん依存関係グラフの奥深くに組み込まれると、それはパッチのサイクルよりも長く生き残ります。 AIワークロードはいま、この力学を引き継いでおり、依存関係が常に最新であることを前提としたリリース頻度がそれをさらに悪化させています。

AIパッケージで最も多く見られるCVEトップ10

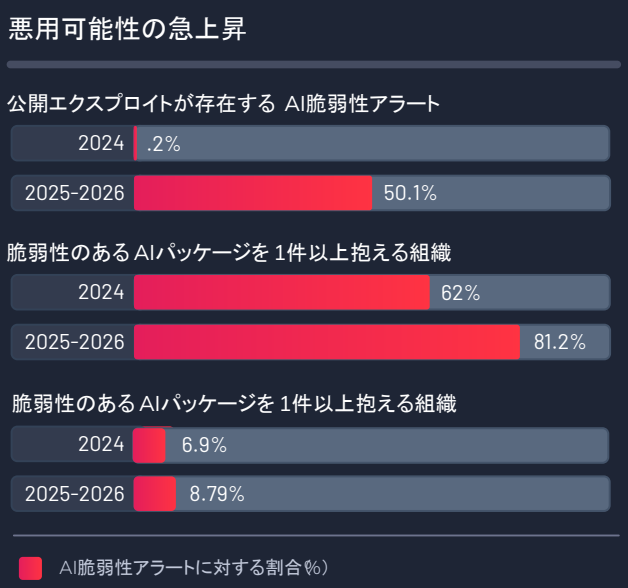
順位	CVE	パッケージ	AI導入企業に対する割合（%）	CVSS	深刻度
1	CVE-2026-25990	Pillow	72.8%	8.9	HIGH
2	CVE-2021-34141	numpy	46.9%	5.3	MEDIUM
3	CVE-2024-5206	scikit-learn	32.9%	5.3	MEDIUM
4	CVE-2025-6985	Pillow	28.8%	7,5	HIGH
5	CVE-2024-28219	Pillow	27.8%	6.7	MEDIUM
6	CVE-2025-6984	langchain-community	25.8%	7.5	HIGH
7	CVE-2025-68665	langchain-core	21.3%	9.1	CRITICAL
8	CVE-2024-43598	lightgbm	21.2%	8.1	HIGH
9	CVE-2026-33682	streamlit	20.9%	4.8	MEDIUM
10	CVE-2025-69662	geopandas	19.3%	8.6	HIGH

攻撃を受けるAIサプライチェーン

2.3 悪用可能性の急上昇

AIパッケージの脆弱性における悪用可能性は、例外的なケースから当たり前の状態へと変わりました。2024年、公開エクスプロイトが存在する AI脆弱性アラートは 0.2%でした。本レポートではその数字が50.1%、2年で250倍の増加です。

2024年には、悪用可能性の低さが AIパッケージの修正を後回しにする理由として挙げられていました。その言い分はもはや通りません。それにもかかわらず、修正が提供されている AI脆弱性アラートの99.9%が未適用のままです。今日にでも解消できる 104,390件の固有アラートのうち、104,284件が未対応で残っています。攻撃者は公表済み AIパッケージ脆弱性の半数について実際に動くエクスプロイトを持ち、防御側は利用できる修正の約 1,000件に1件しか適用していません。AIのリリースサイクルは、それが依存するパッチのサイクルより速く回っており、その差は広がり続けています。



CVSS平均スコアが高いAIパッケージトップ10

順位	パッケージ	CVSS平均	AI導入企業に対する割合（%）
1	mlflow	9.95	**30.2%**
2	tensorflow	9.76	**36.0%**
3	vllm	9.70	**8.7%**
4	keras	9.61	**37.0%**
5	PyTorch	9.56	**42.5%**
6	ray	9.26	**37.8%**
7	gradio	9.08	**16.4%**
8	Pillow	8.87	**92.1%**
9	llama-index	8.86	**14.2%**
10	onnx	8.83	31.8%

攻撃を受けるAIサプライチェーン

2.4 深刻度が最も高いAIパッケージ

本章の前半で示した分布データは、AIパッケージの脆弱性がいかに広く広がっているかを示すものでした。本節では、その脆弱性が最も影響の大きいパッケージにおいてどれほど深刻かを見ていきます。トップ10のすべてがCVSS平均8.83以上で、うち8つは9.0以上です。AIの基盤スタックは、初期状態からほぼ最大級の深刻度で稼働しているのです。

最も重大な脆弱性は、AIスタックの基盤レイヤーに存在します。9.95のMLflowは多くのMLチームにとって事実上のモデルレジストリであり、そのほぼ最大値に近いスコアは、信頼された社内利用を前提に設計されたパッケージが、いまや企業ネットワーク上に置かれている現実を映しています。TensorFlow、Keras、PyTorchは、モデルそのものの学習と展開に使われるフレームワークです。9.70のvLLMは、オープンウェイトの本番環境の多くを支える推論レイヤーで、現時点の規模は小さいものの急速に伸びています。9.26のRayは、公開状態のクラスタが数百規模で侵害された2024年のShadowRayキャンペーンの標的となったパッケージです。

8位のPillowは、トップ10のなかで唯一、ほぼ全面的に普及していながら（AI利用組織の92.1%）、Criticalに近い深刻度（8.87）を併せ持つパッケージです。

総合すると、この表を読むCISOは、これらのパッケージのうち少なくとも数個が自社環境に存在し、その多くがほぼ最大級の深刻度にあると想定すべきです。AI基盤スタックの修正頻度は、一般的なSLAの枠内に収めておくことはできません。

攻撃を受けるAIサプライチェーン

2.5 新興AIツールの脆弱性

Orca Research Podは、前回レポートの時点では存在しなかったパッケージについても脆弱性のシグナルを追跡しています。これらのツールは、脆弱性の傾向が把握されるより速く企業環境に到達しており、初期データはどこに露出が形成されつつあるかを示しています。 2026年のテレメトリでは 3つのカテゴリが浮かび上がりました。ホスト型モデルを呼び出すために使われる SDK、エージェントや連携をオーケストレーションするフレームワーク、そして Model Context Protocol（MCP）のエコシステムです。MCPは近年、主要な 3パッケージで成長が加速しているカテゴリです。

エージェント／連携フレームワーク

このカテゴリには、エージェントのオーケストレーションや既存インフラとモデルの接続に使われるライブラリが含まれます。 678件のアラートを持つ langchain-openaiは、本節において MCP以外で最大の単一露出です。

ベンダーSDK

パッケージ	アラート件数	備考
google-cloud-aiplatform	376	Google Vertex AIのSDK。AnthropicとIBMを合わせたよりも大きな導入規模
@ibm-cloud/watsonx-ai	30	IBM watsonx.aiのSDK
anthropic（Python）	21	アプリケーションコードからAnthropicのモデルを呼び出すための本番用SDK
@anthropic-ai/sdk（TypeScript）	16	Anthropicモデル向けTypeScript SDK
@anthropic-ai/claude-code	6	インストール型のClaude Codeパッケージ。開発者のワークステーション上で動作
@github/copilot	5	GitHub CopilotのSDK

エージェント／連携フレームワーク

パッケージ	アラート件数	備考
langchain-openai	678	LangChainのOpenAI連携。MCP以外では新興ツール中で最大の露出
pydantic-ai	49	新興のエージェントオーケストレーションフレームワーク。LLMエージェント向けの構造化出力レイヤー
Crawl4AI	2	AIに最適化されたWebスクレイピングフレームワーク

MCPエコシステム

パッケージ	アラート件数	備考
mcp（Python SDK）	425	Model Context Protocolのリファレンス実装となるPython SDK
@modelcontextprotocol/sdk（TypeScript）	365	リファレンス実装のTypeScript SDK
fastmcp	341	高性能なMCPサーバーフレームワーク
github.com/modelcontextprotocol/go-sdk	9	Go SDK
MCP Inspector（CVE-2025-49596）	9	MCPサーバーのデバッガにおけるRCE（CVSS 9.4）
io.modelcontextprotocol.sdk:mcp-core（Java）	13	Java SDK
mcp-atlassian	2	Atlassian製品向けMCPサーバー

MCPエコシステム

Model Context Protocolは2024年末に公開され、いまでは大規模に本番稼働しています。3つのリファレンス SDK（Python、TypeScript、fastmcp）だけで1,100件を超えるアラートが、わずか 18か月足らずのリリース履歴に集中しています。 MCPは2024年版レポートには登場していませんでしたが、いまや Orca Research Podが追跡する新興ツールのなかで最大のカテゴリです。

当社テレメトリでは、mcp-handler、mcp-framework、mcp-server-git、mcp-run-python、a11y-mcpなど、より小規模な MCPサーバーも確認されました。いずれも上記の表に掲載する基準には達していません。

これら3つの表から、2つの傾向が際立ちます。第一に、 MCPはもはや「数が小さい話」ではありません。すでに 3つのパッケージが大手ハイパースケーラーのベンダーSDKと同水準の規模に達し、このエコシステムは公開から 18か月のうちに CVSS 9.4のRCEを生み出しました。第二に、エージェント／連携カテゴリで最大の単一露出はlangchain-openai（678件のアラート）であり、その規模は Vertex AIのSDKに匹敵し、Anthropic、IBM、GitHub CopilotのSDKを合計した数を上回ります。攻撃対象領域は導入速度が最も速い地点に形成されつつあり、 MCPの規模ではもはや静かに形成されているとは言えません。これらのパッケージは、本章の前半で示した基盤 AIスタックと同等の扱いで、SBOMとパッチSLAの対象に含めるべきです。

03 AIエージェントとRAG：ガバナンスなき攻撃対象領域

3. AIエージェントとRAG：ガバナンスなき攻撃対象領域

AIエージェントは実際のクラウド権限を持って動作し、 RAGパイプラインを通じて企業データにアクセスし、自律的に行動します。 2025年から2026年にかけて起きた一連のインシデントは、こうしたシステムに安全制御が欠けている場合、そしてエージェントが安全確保に必要なガバナンスの枠組みより速く配備された場合に何が起きるかを示しました。

3.1 エージェントが暴走するとき

その5か月後の2026年2月、セキュリティ研究者の Adnan Khan氏が「Clinejection」を公開しました。これは、コーディングアシスタント Clineが備えるGitHub Issueのトリアージボットを、メンテナー自身に対する武器として転用したプロンプトインジェクション攻撃です。攻撃の連鎖は、間接的なプロンプトインジェクション、 GitHub Actionsのキャッシュポイズニング、認証情報モデルの弱点を 1つのエクスプロイトに束ね、公開GitHub Issueを開くだけで発動しました。さらに、正体不明の攻撃者が同じ不備を利用して、許可されていない cline@2.3.0パッケージを npmに公開し、8時間のうちに約4,000台の開発者マシンへ OpenClaw AIエージェントをインストールさせました。Clineの利用者は500万人を超えます。

2026年のSANS「State of Identity Threats and Defenses」調査では、非人間アイデンティティが76%増加し、74%の組織が認証情報を必要とするAIエージェントを既に利用していること92%がマシン認証情報を90日以内にローテーションできていないこと、そしてセキュリティリーダーの70%が、自社環境でエージェント型AIが稼働しているかどうかすら把握していないことが報告されました。

ガバナンス側の対応は進んでいます。 OWASPは2025年12月にエージェント型アプリケーション向けの Top 10を公開し、エージェント固有のリスクに関する初の公式な分類を確立しました。 Microsoftは2026年4月にAgent Governance Toolkit（https://github.com/microsoft/agent-governance-toolkit/）をリリースしており、OWASPのAgentic Top 10の全10項目を網羅すると謳うオープンソースのランタイムガバナンスレイヤーです。

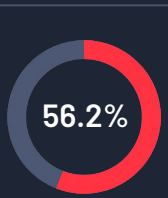
要するに、ガバナンス側の対応は前進しているものの、導入側がそれに追いついていないということです。以下に示す調査結果は、これらのインシデントやフレームワークが指し示すコントロールを実装できている組織がいかに少ないか、そしてエージェントのどれほど多くが、デフォルトの権限、デフォルトのログ設定のまま、エージェントの動作と本番システムの間に実行時の境界を持たずに配備されているかを示しています。

3.2 エージェントフレームワークの導入状況

AI導入企業の 56.2%が、本番環境でエージェントフレームワークを稼働させています。首位のLangGraphは組織の 25%超で導入され、次点の AutoGenは3%で、8.5倍の差がついています。エージェントフレームワークの層は、 AIスタックのどのカテゴリよりも速く単一プロジェクトへ集約されつつあります。その結果、 LangGraphの依存関係グラフは、今日の本番 AIにおいて最も集中したサプライチェーンリスクの 1つとなっています。

AWSでは、Bedrockを利用する組織あたり平均で約 15のエージェントがクラウド権限を持って稼働しています。それにもかかわらず、Bedrock利用組織の57.1%はガードレールを設定していません。

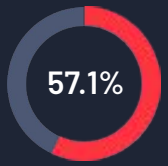
Orca Research Podによる「AI-Induced Lateral Movement（AILM）」の研究は、統制されないエージェントがなぜ危険なのかを示しています。AIエージェントが後に処理するクラウドリソースのタグやCRMのフィールドにプロンプトを注入することで、攻撃者はエージェントを武器化してコマンドを実行させ、AIレイヤーを通じて水平移動できるのです。クラウド権限を持ち、かつガードレールがないエージェントは、AILMにとって理想的な条件を提供します。



AI導入企業の 56.2%が、本番環境でエージェントフレームワークを稼働



AI導入企業の 25%が、LangGraphを導入



Bedrock利用組織の 57.1%が、ガードレールを未設定

3.3 RAG：AIと企業データの接続

AI導入企業の 64%がベクターデータベースを導入し、 LLMを社内文書、顧客記録、独自ナレッジに接続しています。その市場は大きく分散しています。

7種類のベクターデータベースが、それぞれ 25%を超える導入率に達しています。 RAGを運用する組織は平均3.78種類のベクターデータベースを同時に使っており、つまり一貫したセキュリティポリシーを、複数のプラットフォーム、複数の提供形態（マネージドと自社ホスト）、複数のアクセスパターンにまたがって適用しなければなりません。これは攻撃者の存在を考慮する前の、運用上の前提条件です。

AWSでは、Bedrockを利用する組織あたり平均で約 21件の承認済み Bedrock Knowledge Baseデータパイプラインが稼働しています。その一つ一つが、 LLMと膨大な社内文書、顧客記録、独自ナレッジとを結ぶ生きた接続です。なお、この規模には Bedrock外で構築された非承認の RAGパイプラインは含まれていません。AIスプロールの章で示す可視性のデータからすると、その数はさらに大きい可能性があります。

研究はこの脅威モデルを裏付けています。 RefineRAG13は、汚染文書を用いて標準的な RAGベンチマークで90%の攻撃成功率を達成しました。しかも、その文書は検証した攻撃手法のなかで文法エラー率と反復率が最も低く、人間のレビュアーにも自然に読めるということです。 BadSkill14はバックドアを仕込んだエージェントスキルで 99.5%の攻撃成功率を達成し、汚染率 3%でも91.7%に達しました。この脅威モデルが狙うのは、まさに LangGraphやAutoGenが企業内に広めているインストール型スキルのパターンです。また Semantic Intent Fragmentation15は、一見正当に見える 1つの要求を、個別には無 害なサブタスクへ分解した場合、 GPT-20Bのオーケストレーターが 71%のケースでポリシー違反の計画を生成することを示しました。これらの攻撃はいずれも、当社テレメトリで数百社が現に本番運用している構成を標的としており、その対象は AI導入企業の 56.2%が稼働させているマルチエージェントのパイプラインです。



13. Wang et al., "RefineRAG: Word-Level Poisoning Attacks via Retriever-Guided Text Refinement," arXiv:2604.07403, 2026年4月
14. Tie et al., "BadSkill: Backdoor Attacks on Agent Skills via Model-in-Skill Poisoning," arXiv:2604.09378, 2026年4月10日
15. Ahad et al., "Semantic Intent Fragmentation: A Single-Shot Compositional Attack on Multi-Agent AI Pipelines," arXiv:2604.08608, 2026年4月8日（2026年4月20日改訂、AAAI 2026 Summer Symposium採択）



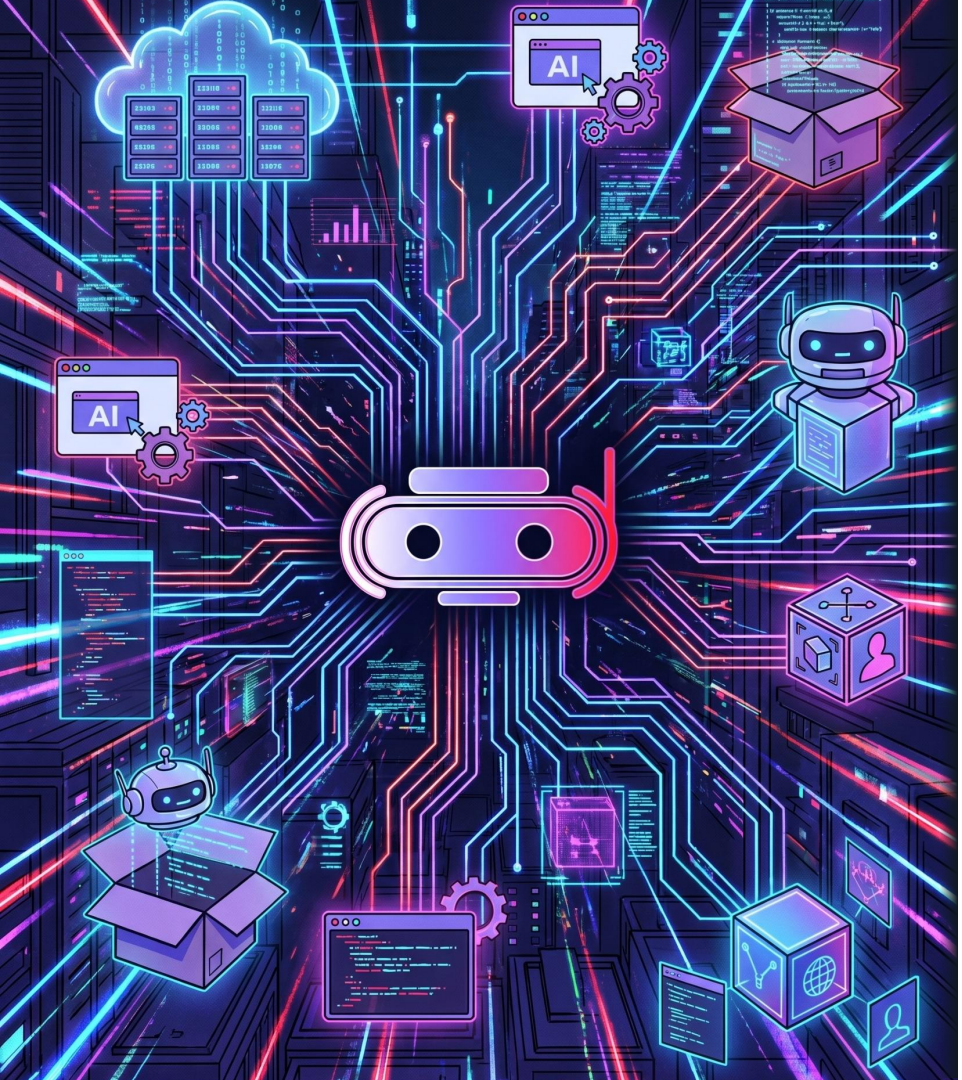
攻撃を受けるAIサプライチェーン

3.4 ガードレールのギャップ

エージェントはクラウド権限を持ち、 RAGを通じて企業データに接続し、その半数以上が安全制御なしで動作しています。Orcaのテレメトリでは、許可されていない Bedrock GuardrailのIAMロールがAI関連アラートのなかで最も多い件数を占めており、しかもそうした組織の大半は Bedrockそのものを導入していません。これは、過度に広範な IAMポリシーによって、AWS環境全体で不要な AIサービス権限が付与されていることを示しています。

ただし、ツールは揃っています。クラウドネイティブな適用には Bedrock Guardrails、ランタイムのポリシー適用にはMicrosoftのAgent Governance Toolkit、分類の網羅にはOWASPのTop 10 for LLM ApplicationsとAgentic Top 10があります。本章全体の調査結果が示すのは、これらのツールが統制すべきエージェントやパイプラインの導入ペースに対して、ツール自体を採用できている組織がいかに少ないかということです。 AI導入企業の 56.2%が本番でエージェントフレームワークを稼働させ、64%が企業データに接続されたベクターデータベースを持ち、 Bedrock利用組織の 57.1%はプラットフォーム標準のガードレールを設定せずにエージェントを運用しています。エージェントの導入とエージェントのガバナンスの差は広がり続けており、続く各章のデータは、その差がアイデンティティ、インフラ、暗号化の状況にどう現れているかを示します。

04 AIスプロールとガバナンスのギャップ



4. AIスプロールとガバナンスのギャップ

AIは、統制すべき単一のサービスではありません。モデル、エージェント、パッケージ、ブラウザ拡張機能、クラウドサービスが入り組んだ広がりであり、それはセキュリティチームが把握できる速度を超えて、多くの企業に浸透しています。前章までで、その構成要素を挙げてきました。脆弱性を抱えた AIパッケージ、クラウド権限を持ちながらガードレールがほとんどないエージェントフレームワーク、企業データに接続されたベクターデータベース、そして AWS、Microsoft、OWASPが提供してきたコントロールが、エージェントの導入ペースに見合うほど採用されていないことを示すアイデンティティ層の調査結果です。本章では、その総体を示します。 Orca Research Podが組織データセット全体のテレメトリから見出したのは、 AIを導入している組織が平均して 4種類以上の異なるAIサービスカテゴリを同時に運用しており、シャドー AIの広がりがそのギャップを四半期ごとに拡大させているという事実です。

AIスプロールとガバナンスのギャップ

4.1 AIの可視性ギャップ

2026年4月、CSO Onlineは、AIの導入がガバナンスを追い越すなかでセキュリティリーダーが可視性の維持に苦慮していると報じました。 AI機能へのアクセスが速く容易であるということは、セキュリティチームがその存在を知る前に、クラウドアカウント、開発者のマシン、業務システムにワークロードが現れるということです。金銭的な影響も理論上の話ではありません。 IBMの「2025 Cost of a Data Breach Report」では、シャドー AIが関与した侵害は、関与しない侵害に比べて平均 67万ドル高くつき、侵害を受けた組織の 20%が、侵害の少なくとも一部の原因はシャドー AIの関与にあるとしています。

シャドーAIは問題をさらに複雑にします。従業員はセキュリティレビューを経ずに AIツール、ブラウザ拡張機能、API型サービスを組み込みます。LayerXの「2026 Browser Extension Security Report」によれば、AIツールはセキュリティチームが監視する境界の外側で動作することが多く、しかも利用者が社内のシングルサインオン、顧客データ、社内ドキュメントにアクセスするのと同じブラウザセッション上に存在しています。

$670K

シャドーAIが関与した侵害は、関与しない侵害に比べて平均 67万ドル高いコストが発生



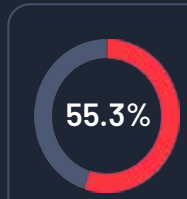
侵害を受けた組織の 20%が、侵害の少なくとも一部の原因をシャドーAIの関与に帰しています。

AIスプロールとガバナンスのギャップ

4.2 AIサービスのスプロール

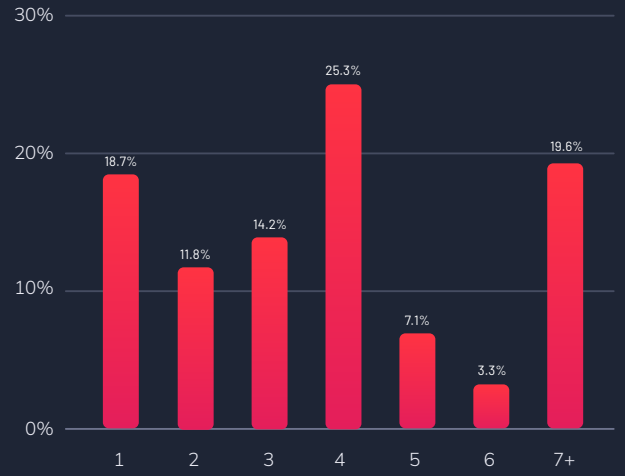
AIクラウドサービス利用者の 55.3%が4種類以上の異なる AIサービスを、19.6%が7種類以上を運用しています。追跡対象の 18種類すべてを利用している組織も 3社ありました。最も多い構成は4種類で、AIクラウド利用者の 25.3%を占めます。

各サービスは、それぞれ独自のセキュリティモデル、暗号化オプション、アクセス制御、コンプライアンス要件を持ちます。SageMaker、Azure OpenAI、Vertex AI、Bedrockは、いずれも異なるデフォルト設定、異なるアイデンティティモデル、異なるネットワーク分離の考え方で提供されています。これらを4つ同時に、しかも多くの場合は複数のクラウドプロバイダーにまたがって統制することが、CISOが訴える可視性ギャップの中核的な難しさです。 6種類以上を運用する組織は同じ問題を掛け算された複雑さで抱えることになり、 18種類すべてを運用している 3社は、どの単一ベンダーのドキュメントでも網羅できない状態で、統合された AIセキュリティ基準を持たずに運用していることになります。



AIクラウドサービス利用者の 55.3%が、4種類以上の異なるAIサービスを運用

クラウド利用者あたりのAIサービス種類数



4.3 本番環境におけるAIコード生成

AI導入企業の 33%がGitHub Copilotを導入しています。 Copilot以外にも、2025年から2026年にかけて AI IDEのエコシステムは急速に広がり、 Cursor、Windsurf、Claude Code、Amazon Q Developer、Gemini Code Assistがいずれも支持を集めました。

セキュリティ上の課題は、ツール自体にとどまりません。 Orca Research Podによる脆弱性研究「RoguePilot」は、AIコーディングツールがそれ自体で脆弱性を持ち込み得ることを示しました。しかし、より本質的なガバナンスの問題は運用面にあります。

組織はいま、人間のレビュアーが到底追いつけない速度で AIが生成するコミットに対して、コードレビュー基準、シークレットの適切な扱い、セキュリティポリシーを適用しなければなりません。AIが生成したコードは、その組織のセキュリティ上の作法に関する組織知を引き継ぎません。そして、従来のレビュープロセスが想定していなかった種類の脆弱性を持ち込む可能性があります。

2023年に公開されたスタンフォード大学の研究 16では、AIコーディングアシスタントを使う開発者は、使わない開発者に比べて明確に安全性の低いコードを書く一方で、自分のコードを実際よりも安全だと評価することが分かりました。脆弱性の発生率が高まると同時に開発者の自信も高まるというこの組み合わせが、いま AI IDEのエコシステムが中心に位置する構造的リスクです。

4.4 規制上の露出

当社テレメトリでは、 AIクラウドサービス利用者の 19.1%がEUリージョンで AIサービスを稼働させています。 EU AI Actの禁止行為に関する規定は 2025年2月に、GPAI（汎用AI）に関する義務は 2025年8月に発効しました。高リスクシステムに関する義務は 2026年8月2日に発効し、罰金は最大 3,500万ユーロ、または全世界年間売上高の 7%に達します。採用選考、与信判断、必須サービスへのアクセス、重要インフラで AIを稼働させている組織は、自社システムが高リスク分類に該当するかどうかを、いま評価しておくべきです。

米国では、連邦レベルの姿はまだ形成途上です。トランプ政権が 2026年3月に示した国家政策フレームワークは、州の AI法を連邦法で先占し、監督は既存の分野別当局に委ねるよう議会に求めています。州レベルでは、コロラド州が2024年のAI法を施行前に改正し（S.B. 26-189）、雇用、金融サービス、医療、住宅における重大な意思決定に実質的な影響を与える自動意思決定技術に対象を絞った、より限定的な枠組みに置き換えました。同法は 2027年1月1日に施行され、施行規則はまだ策定中です。罰則は違反 1件あたり最大 2万ドルで、執行はコロラド州司法長官が独占的に担います。

中国は、世界で最も詳細な AI規制体制を引き続き前進させています。 2026年1月1日に施行された改正サイバーセキュリティ法には AIコンプライアンスに関する専用規定が追加され、AI生成物のコンテンツ表示に関する規則は 2025年9月から施行されています。

16. Perry et al., "Do Users Write More Insecure Code with AI Assistants?," ACM CCS 2023, arXiv:2211.03622

05 AIの認証情報と安全でないアクセス

5. AIの認証情報と安全でないアクセス

AIの導入は、シークレット露出の新しいカテゴリを生み出しました。 AIサービスの API キーは、独自モデルへのアクセス、 RAGパイプラインを通じた機密の企業データへのアクセス、そして高額な GPUや推論リソースの従量課金へのアクセスを与えます。 AI キーが1つ侵害されるだけで、知的財産の窃取、データの持ち出し、そして大規模な不正クラウド利用が数時間のうちに起こり得ます。この攻撃類型は 2024年に「LLMjacking」として記録され、観測された事例では被害者 1社あたりの1日の請求額が4万6,000ドルに達しました。金銭的被害は、そのキーがアクセスできるモデル、データセット、下流システムの数だけ膨らんでいきます。

当社テレメトリでは、AI導入企業のおよそ 10社に3社（29.5%）が、AIのシークレットまたはキーを少なくとも 1つ安全でない場所に保管しています。プロバイダー別の母集団には重複があります。AI導入企業は平均して 2つのAIプロバイダーと同時に連携しており、認証情報が漏れ得る面積はその分だけ広がります。

gitの履歴にコミットされたキーは、現在のコードベースから削除された後も復元可能なまま残ります。コミット履歴に含まれる認証情報は、リポジトリにアクセスできる者なら誰でも抽出でき、AIサービスへの継続的なアクセスを与えてしまいます。公開リポジトリは自動化された認証情報収集ツールによって常時スクレイピングされており、ローテーションを行っても、すでに公開されたものを取り消すことはできません。 OpenAIや Hugging Faceのキーが公開コミットで一度でも見える状態になったなら、少なくとも 1つの第三者に収集済みだと考えるのが唯一安全な前提です。



06 AIインフラの露出

6. AIインフラの露出

AIワークロードは安全でないデフォルト設定のまま提供され、組織がそれを変更することはほとんどありません。最も広く導入されているIクラウドサービスが、同時に最も露出しているサービスでもあり、脅威インテリジェンス各社が025年から2026年にかけて警告してきたAIインフラへの攻撃経路が、あらかじめ用意された状態になっています。

6.1 標的とされているAIインフラ

CrowdStrikeの「2026 Global Threat Report」は、AIを活用する攻撃者による攻撃が前年比89%増加したことを記録しています。 eCrimeの平均ブレイクアウト時間、すなわち初期侵入から 2台目のシステムへ水平移動するまでの時間は 29分に短縮され、2024年から65%も加速しました。観測された最速のブレイクアウトは 27秒でした。ある侵入では、初期侵入から 4分以内にデータの持ち出しが始まっています。攻撃者はいま、18か月前には現実的でなかった速度で動いており、 AIインフラはますます魅力的な標的になっています。

2026年4月、セキュリティ研究者が Google CloudのVertex AI Agent EngineおよびAgent Development Kitにおける一連の脆弱性を公開しました。 Per-Project, Per-Product Service Agent（P4SA）に付与された過大なデフォルト権限により、攻撃者は侵害した AIエージェントから GCPのサービス認証情報を抽出し、所有者のプロジェクトやデータストレージへ横展開できました。エージェントを内部脅威に変えてしまう構図です。研究者はさらに、エージェント環境内でリモートコード実行に悪用され得るファイルパスも特定しており、永続的なバックドアへの経路が開いていました。 Googleはこの報告を受けてドキュメントを更新し、最小権限で実行するために Bring Your Own Service Accountの利用を推奨しています。

同じ月、Noma Securityは「GrafanaGhost」を公開しました。これは、認証もユーザー操作も不要で GrafanaのAIガードレールを回避する間接プロンプトインジェクション攻撃です。攻撃は被害組織の外部から与えられる URLクエリパラメータを使い、GrafanaのAIが処理する隠し命令を仕込みます。特定のトリガーキーワードによって、モデルは自身の安全制約を無視してしまいます。データの持ち出しは AIが自ら開始する Markdown画像の処理を通じて行われ、観測者からは正当な AIの動作に見えるため、従来の外向き通信の監視をすり抜けます。Grafanaは責任ある開示を受けて修正を提供しました。

これら3つの事例は、2026年における AIインフラに対する攻撃手法の変化を示しています。計算資源の窃取は依然として動機の 1つですが、いまはより運用上危険なパターンとして、水平移動、認証情報の抽出、そして AIが起点となる経路を通じた静かなデータ持ち出しが挙げられます。これらは、多くの組織がいまも頼っている外向き通信の制御や検知の仕組みをすり抜けてしまいます。

AIインフラの露出

6.2 SageMakerの攻撃チェーン

Amazon SageMakerは、最も広く導入されている AIクラウドサービスです（全組織の 19.5%）。そして、SageMaker利用組織の 80%が、中核となる 5つのデフォルト設定をすべて有効のまま運用しています。各デフォルト設定は単独でも設定不備ですが、組み合わさると完全な攻撃経路を形づくります。

バケット命名に関する 92.9%という結果は、本節で最も対処しやすい項目です。旧来のデフォルトパターンである sagemaker-{region}-{account_id}は完全に公開されており、機械的に列挙できます。AWSのアカウント IDは秘密情報ではなく、リージョンも公開 DNSやASNの記録から推測可能です。AWSがデフォルトの命名規則にランダムな文字を追加した後も、 SageMaker利用組織の大多数、修正後にバケットを作成した組織も含めて、推測可能な名前を使い続けています。

残りのステップは、1本の連鎖につながります。ノートブックのインターネット直接アクセスが侵入口となり、root権限とIMDSv2の欠如が、AWSインスタンスメタデータサービス経由での権限昇格と認証情報の窃取を可能にします。これは、認証の欠如が認証情報の抽出を許した 2019年のCapital One侵害と同じコンポーネントです。 VPCによる分離がなければ水平移動は制限されず、暗号化の欠如は保存データを露出させたままにします。ノートブックに管理者権限の IAMを付与している 6.3%の組織では、この連鎖はクラウドアカウントの完全な乗っ取りで終わります。本レポートの前半で扱ったエージェントのガバナンスと認証情報の露出に関する調査結果が示す、最悪の経路です。

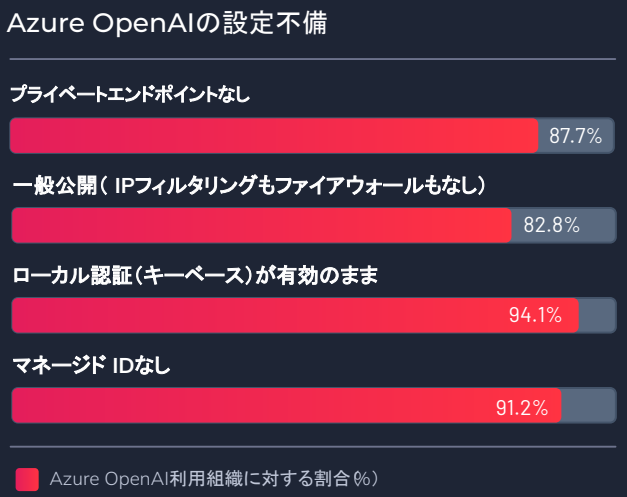


AIインフラの露出

6.3 Azure OpenAI：デフォルトで一般公開

Azure OpenAIは、当社テレメトリにおいて Azure顧客の間で2番目に導入されている AIクラウドサービスであり、監視対象の Azure利用組織の 50.5%が導入しています。そして、これらの環境の大多数はアクセス制御がほとんどないまま一般公開されています。

3つの露出が組み合わさって 1本の攻撃経路になります。Azure OpenAI環境の83%が一般公開され、94%がキーベース認証に依存し、91%はマネージド IDを設定していません。APIキーを盗んだ者は誰でも、その組織の AIモデル、推論エンドポイント、そしてそこで処理されるあらゆるデータに無制限にアクセスできます。本レポートの前半で示した認証情報の露出データでは、 OpenAI利用者の 28.4%が少なくとも1つのキーを露出させています。つまり、盗まれたキーによる Azure OpenAI環境への攻撃は、理論上の懸念ではありません。Microsoftは数年前からキーベース認証を非推奨とし、 Entra IDへの移行を促してきました。当社テレメトリにおけるローカル認証 94%という数字は、その非推奨の指針が運用上の移行につながっていないことを示しています。



AIインフラの露出

6.4 Vertex AI：武器化可能で、暗号化されていない

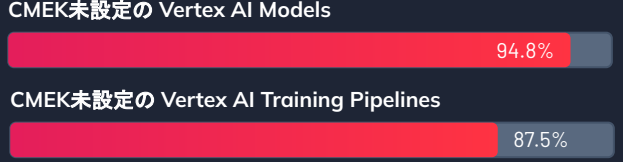
Google Vertex AIは、Palo Alto Networksが武器化可能であることを実証したプラットフォームです。同時に、自社で暗号鍵を管理している組織がほぼ存在しないプラットフォームでもあり、カスタマー管理暗号鍵（CMEK）が設定されているのは 12件に1件未満です。当社テレメトリにおける Vertex AIの母集団はSageMakerやAzure OpenAIより小さいものの、モデルの保存領域と学習パイプラインの双方で暗号化の状況が一貫していることから、この傾向はサンプル上の偶然ではなく、プラットフォーム標準の挙動を反映していると考えられます。

Vertex AI利用組織の 92.9%17が、保存データに対するカスタマー管理の暗号化を設定していません。2024年の98%からは小幅に改善しました。学習パイプライン（ 87.5%がCMEK未設定）とモデルの保存領域（ 94.8%）はいずれも、圧倒的にプロバイダー管理のデフォルト暗号化に依存したままで、学習データ、モデルの重み、推論結果に誰がアクセスできるかを組織側で制御する余地が限られています。

本章で取り上げた 3つのクラウド AIプラットフォームすべてで、同じ運用上のパターンが繰り返されています。プロバイダーが提供する初期設定は組織が必要とするセキュリティ水準に合っておらず、そのギャップを埋められている組織はごく少数だということです。

17. Vertex AI ModelsおよびTraining Pipelinesの加重平均

Vertex AIの暗号化状況



07 AIの暗号化：欠落したレイヤー



7. AIの暗号化：欠落したレイヤー

AIワークロードは、組織のなかでも最も機密性の高いデータを処理します。独自の学習データセット、数百万ドル規模の価値を持つモデルの重み、そして規制対象の顧客情報を含むことも多い推論の入力と出力です。それにもかかわらず、主要　3クラウドプロバイダーのいずれにおいても、大多数の組織はプロバイダー管理のデフォルト暗号化に完全に依存しており、そのデータに誰がアクセスできるのか、鍵はいつローテーションされるのか、どのような監査ログが得られるのかについての制御を手放しています。本章では、AWS、Azure、Google Cloudに共通して見られ、2024年版「State of AI Security」レポート以降ほとんど改善していない、暗号化体制のギャップを示します。

AIの暗号化：欠落したレイヤー

7.1 AIにおいて暗号化が重要な理由

カスタマー管理の暗号鍵がなければ、組織は自社の AI データに誰がアクセスしたかを独立して制御・監査することができません。プロバイダー管理の鍵でも保存データは暗号化されますが、鍵のライフサイクルを制御しているのは顧客ではなくクラウドプロバイダーです。つまり組織は、独自にアクセスを取り消すこともできず、自社の鍵ローテーション方針を強制することもできず、鍵の利用状況に対する可視性も限られます。独自の学習データ、規制対象の顧客情報、価値の高いモデルの重みを扱う AIワークロードにおいて、カスタマー管理の暗号化は、自社データとプロバイダーの間に顧客自身を置くことができる唯一のコントロールです。しかし、大多数の組織はそれを設定していません。

7.2 クラウドを横断した暗号化のギャップ

主要3クラウドプロバイダーで AIサービスを利用する組織のうち、87%から98%がカスタマー管理暗号鍵を設定していません。これは業界全体に共通する傾向です。 AWS、Azure、Google Cloudの顧客は、いずれも同程度の比率で暗号化をプロバイダー任せにしています。

AIサービス別の暗号化状況

クラウドプロバイダー	確認対象のAIサービス	CMEK未設定の平均割合（%）	最大のギャップ
AWS	SageMaker Notebooks、Bedrock Models、Bedrock Training Data、SageMaker Endpoints	71.4%	SageMaker Notebooks：98.4%
Azure	Azure OpenAI Accounts	91.6%	Azure OpenAI Accounts：91.6%
GCP	Vertex AI Training Pipelines、Vertex AI Models	91.2%	Azure OpenAI Accounts：91.6%

7.3 AIサービス別の暗号化状況

特筆すべき結果は3つあります。

最も深刻なのは SageMaker Notebooksの98.4%です。SageMakerのノートブックを稼働させているほぼすべての組織が、カスタマー管理の暗号化なしで運用しており、これはすべての AIサービス、すべてのクラウドを通じて最大のギャップです。

Bedrockのカスタムモデルと、その学習データのバケットが同一の 89.5%というギャップを示しているのは、両者が同じ組織であることを意味します。 Bedrock環境でモデルにカスタマー管理の暗号化を設定していない場合、同じ設定不備がそこに供給される学習データにも波及します。パイプラインが一つの単位として扱われるため、たった 1つの設定判断がML（機械学習）のライフサイクル全体を露出させてしまうのです。

例外は SageMaker Endpoint Configsの8.3%です。AIワークロードが本番トラフィックを処理する場面では、組織は高い比率でカスタマー管理の暗号化を設定しています。ギャップが集中しているのは学習環境と実験環境であり、そこには最も機密性の高い生データ、すなわち独自の学習データセット、モデルの重み、ファインチューニングの成果物、中間出力が実際に存在します。多くの組織は、セキュリティの観点で学習環境を本番同等として扱えていません。しかし、そこにあるデータは本番の推論トラフィックと少なくとも同等の価値があり、多くの場合はより機密性が高いのです。

暗号化は、侵害された後でもデータを守れる唯一のコントロールです。しかし AIワークロードにおいては、それが実質的に不在です。本レポート前半の認証情報の露出に関する調査結果では、プロバイダーによって 25%から40%の組織がAIキーを露出させていました。本章で示した暗号化の状況と組み合わせると、意味するところは明快です。認証情報が侵害された場合、AIの学習データとモデルの重みは、多層防御の最後の受け皿を持たないまま露出するということです。

AIサービス別の暗号化状況

AIサービスおよびリソース種別	CMEK未設定の割合（%）	クラウド
Amazon SageMaker ノートブックインスタンス	98.4%	AWS
Google Vertex AI Models	94.8%	GCP
Azure OpenAI アカウント	91.6%	Azure
Amazon Bedrock カスタムモデル	89.5%	AWS
Amazon Bedrock 学習データバケット	89.5%	AWS
Google Vertex AI 学習パイプライン	87.5%	GCP
Amazon SageMaker エンドポイント設定	8.3%	AWS

08 自社にとって何を意味するのか



8. 自社にとって何を意味するのか

本レポートのデータは、1,200社を超える本番環境の組織が、いままさに AI導入をどう進めているかを切り取ったものです。セキュリティチームは、並行して埋めるべき 2つのギャップに直面しています。1つ目は、AIを本番システムとして扱うことです。稼働中のインフラに当然求められる棚卸し、パッチ適用、認証情報の適切な管理、アクセス制御を、単発の是正作業ではなく継続的に適用すること。2つ目は、従来のセキュリティ機構では想定されていなかったガバナンスのコントロールを立ち上げることです。経営層にとっては、いずれもリソース配分と優先順位付けの判断であり、セキュリティのロードマップに何を加え、本番 AIを担うチームに何を委ねるかを左右します。

当社テレメトリでは半数以上の組織が本番で AIを稼働させていますが、本来であれば本番システムに伴うはずのコントロールは、 AIワークロードの大多数において未成熟なままです。 AIの導入は速く進み、その速度に追いつくガバナンスの枠組みはまだごく限られています。

AIの認証情報の窃取は、被害額が 1日あたり4万6,000ドルに達した事例も記録される、明確に狙われた攻撃類型へと発展しました。 IBMの「2025 Cost of a Data Breach」調査によれば、シャドーAIが関与した侵害には他より 67万ドル高い上乗せコストが生じます。モデルの重みと学習データは、AIスタックのなかで最も機密性の高い層における知的財産の露出そのものですが、SageMaker利用組織の 98%は、カスタマー管理の暗号化を設定せずこれらを危険にさらしています。

一方で、組織が力を集中させた領域では進歩が見え、しかも測定可能です。 Orcaのレポート間の比較では、SageMakerのroot権限は98%から75.8%へ低下し、IMDSv2のギャップは 77%から47.9%へ縮まりました。これらの改善は、 AIワークロードに求められる運用上の規律とあわせて、的を絞った是正が行われた成果です。同じ取り組み方を第 9章のコントロールに当てはめることが、AI時代におけるセキュリティのレジリエンスを高める出発点になります。

09 主要な推奨事項

主要な推奨事項

9.1 直ちに着手すべき対応（0〜30日）

本レポートの調査結果は、一貫した傾向を示しています。私たちが測定したすべての観点で、 AIの導入がAIセキュリティを追い越しているということです。以下の推奨事項は緊急度の順に整理し、本レポートの具体的な調査結果と対応づけています。まず「直ちに着手すべき対応」は、現在の体制のなかで最も悪用されやすいギャップに対処するものです。

1. 修正が提供されている深刻度の高い AIパッケージにパッチを適用する。

当社テレメトリでは、AI導入企業の81.2%が脆弱性のあるAIパッケージを少なくとも1件抱えており、修正が提供されているAI脆弱性アラートの99.9%が未適用のままです。まずはAI利用組織の92.1%に影響するPillowのCVE-2026-25990の是正を優先し、続いて深刻度の高いパッケージ、すなわちmlflow（CVSS平均9.95）、tensorflow（9.76）、keras（9.61）、PyTorch（9.56）へと進めてください。

2. scikit-learnにいま対応する。

scikit-learnは当社テレメトリで最も広く導入されているAIパッケージ（AI導入企業の82.5%）で、影響を受けた組織全体のCVSS平均は7.56です。CVE-2024-5206単体でも、本レポートにおいて3番目に多く見られるAIパッケージのCVEであり、AI導入組織のおよそ3社に1社に存在します。

3. 露出したAIのAPIキーをローテーションし、安全に保管する。

AI導入企業の29.5%が、AIキーを少なくとも1つ安全でない場所に保管しており、少数ながら一定数の組織がgitのコミット履歴にAIキーを残しています。露出したキーはすべて直ちにローテーションし、シークレットはマネージドなシークレットストアに移し、全リポジトリのgit履歴を監査してください。

4. Amazon Bedrockでガードレールを有効にする。

Bedrock利用組織の57.1%が、ガードレールを設定せずにエージェントを稼働させており、許可されていないBedrock GuardrailのIAMロールは当社テレメトリでAI関連アラートの最多件数を生んでいます。本番稼働するすべてのエージェントでAmazon Bedrock Guardrails（コンテンツフィルタリング、根拠の検証、禁止トピック）を有効にし、ガードレール設定を変更または回避できる権限を持つIAMロールを監査してください。

5. SageMakerノートブックの設定を堅牢化する。

SageMaker利用組織の80%が、中核となる5つの安全でないデフォルト設定を同時に有効のまま運用しています。ノートブックインスタンスのroot権限を無効化し、IMDSv2を強制し、機械的に列挙可能な旧来のsagemaker-{region}-{account-id}パターンに従うS3バケットは名称を変更してください。root権限とIMDSv2の変更は、既存インスタンスに対して停止なしで適用できます。

9.2 短期的な取り組み（30〜90日）

1. AIサービス全体でカスタマー管理の暗号化を設定する。

主要3クラウドプロバイダーでAIを利用する組織の87%から98%が、カスタマー管理暗号鍵を設定していません。SageMaker Notebooks（98.4%が未設定）、Vertex AI Models（94.8%）、Azure OpenAI Accounts（91.6%）を優先してください。ギャップは、最も機密性の高いAIデータが存在する学習環境と実験環境に集中しています。

2. AIエージェントの導入状況を棚卸しし、統制する。

AI導入企業の56.2%が本番環境でエージェントフレームワークを導入しており、Bedrock利用組織は平均で約15のエージェントをクラウド権限付きで稼働させています。すべてのエージェントの導入状況を可視化し、その権限を文書化し、最小権限を適用し、エージェントのアイデンティティを人間のアイデンティティと同じ厳格さで扱ってください。

3. ベクターデータベースのアクセス制御を監査する。

AI導入企業の64%がベクターデータベースを導入しており、RAGを利用する組織は平均3.78種類を同時に運用しています。それぞれの環境について、ネットワークの露出、認証要件、データ分類を監査してください。

4. 承認されていない AI利用の可視性を確立する。

AI導入企業の29.5%が少なくとも1つのAIシークレットを露出させており、55.3%が4種類以上の異なるAIサービスを運用しています。AI系ブラウザ拡張機能、承認されていないAPIキー、社用アカウントで動作するAI SaaSツールを検出する仕組みを導入してください。従業員のAI利用を完全に把握できていると回答した組織はわずか30%であり、シャドーAIが関与した侵害は平均で67万ドル高いコストを伴います。

5. EU AI Actおよび州レベルの AI規制への対応準備をいま始める。

AIクラウドサービス利用者の19.1%が、EU AI Actの適用対象となるEUリージョンで稼働させています。高リスクシステムに関する義務は2026年8月2日に発効し、罰金は最大3,500万ユーロ、または全世界売上高の7%です。コロラド州の改正AI法（S.B. 26-189）は2027年1月1日に施行され、それに先立って州司法長官による施行規則が示される予定です。

6. AIが生成したコードに特化したセキュリティレビューのプロセスを整備する。

AI導入企業の33%がGitHub Copilotを導入しており、より広いAI IDEのエコシステム（Cursor、Windsurf、Claude Code、Amazon Q）も急速に広がっています。スタンフォード大学の研究では、AIコーディングアシスタントを使う開発者は明確に安全性の低いコードを書きながら、それをより安全だと評価することが分かりました。コードの作成者が人間かAIかを問わず、すべてのコミットにセキュリティスキャンを必須で適用してください。RogquePilotの研究は、AIコーディングツール自体が攻撃経路になり得ることを示しています。

9.3 戦略的な改善（90日以降）

1. AIエージェントに最小権限のアーキテクチャを採用する。

主モデルが動作する前に、副となるガードレール用モデルが入力と出力を検査するマルチLLM構成を検討してください。また、プロンプトインジェクションによってエージェントが侵害された場合の影響範囲を抑えるため、エージェントの実行環境を分離します。

2. AIに特化したセキュリティ監視を実装する

プロンプトインジェクションの試行、エージェントの異常な振る舞い、許可されていないツール実行、そしてAI-Induced Lateral Movement（AILM：攻撃者がクラウドリソースのタグやRMのフィールドに命令を仕込み、AIエージェントが後にそれを処理する攻撃類型）を検知できるよう計測を組み込んでください。既存のSIEMやEDRでは、これらのパターンは初期状態では捉えられません。

3. RAGのセキュリティ実務を確立する。

ベクターデータベースに取り込むデータは、あらゆる本番システムに入るデータと同じ厳格さで扱ってください。従来のデータパイプラインに存在する出所の確認、検証、監視のコントロールをRAGの取り込みにも同様に適用します。

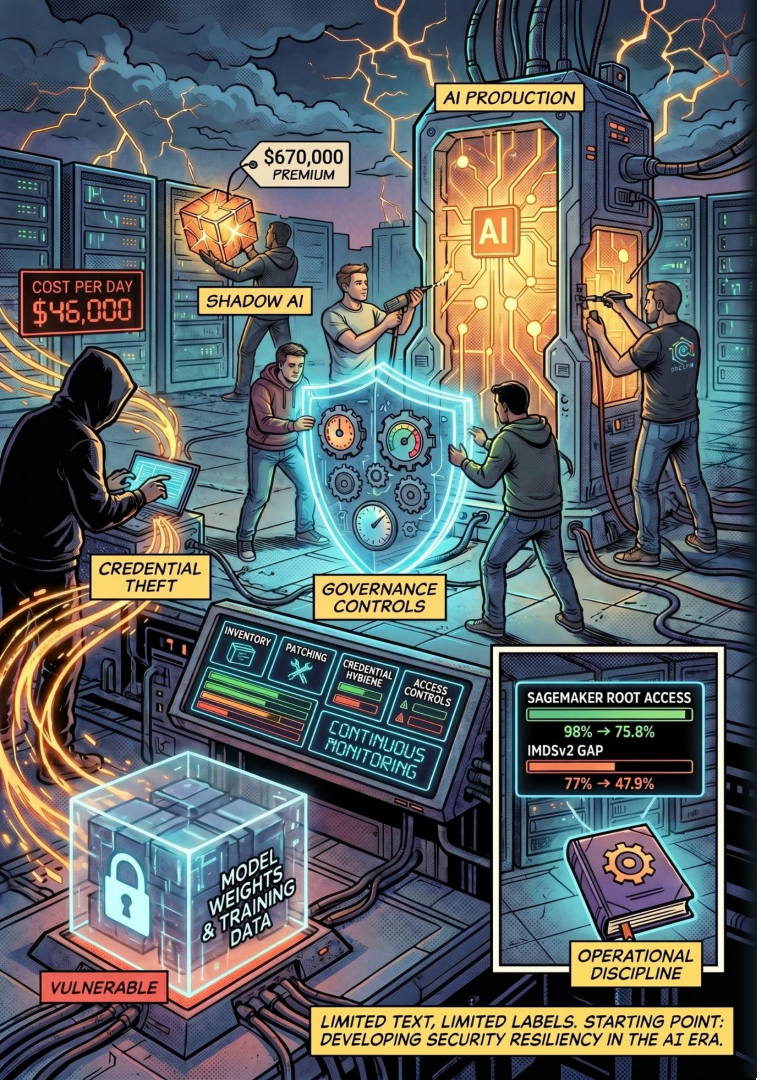
4. クラウドプロバイダーを横断して AIセキュリティのガバナンスを統合する

AIクラウドサービス利用者の35.3%が4種類以上の異なるAIサービスを、19.6%が7種類以上を運用しています。すべてのクラウドを横断したAI資産の統合台帳を整備し、セキュリティ基準を標準化し、暗号化、アクセス制御、ネットワークの露出について一貫したポリシーを適用してください。可能であれば、プロバイダー提供のガードレールを、決定論的なポリシー適用で補完しますMicrosoftのAgent Governance Toolkitのようなフレームワークを使えば、ポリシーをコードとして記述・適用でき、ガバナンス基準をより確実に実装できます。

5. AIスタックのすべての層に本番相当のセキュリティ基準を適用する。

本レポートで最大のギャップはAIの学習環境、実験用ノートブック、エージェントのサンドボックスは、本番環境よりも低いセキュリティ要件やデフォルト設定のままで安全に運用できる、という組織的な思い込みです。実際にはそうではありません。戦略的な投資とは、本番のセキュリティSLA（パッチ適用の頻度、アクセスレビュー、暗号化要件、インシデント対応の範囲）を、「本番」と名付けられているかどうかに関わらず、すべてのAI環境へ拡張することです。

10 まとめ



まとめ

2年前、AIセキュリティは先を見据えた課題でした。いまでは、運用上の現実です。

AIは、安全が確保される速度よりも速く導入されています。半数を超える組織が本番で　AIを稼働させ、81%が脆弱性のあるAIパッケージを運用し、CVSS平均は2024年の6.9から8.79へ上昇しました。そして修正可能なAI脆弱性の99.9%が未適用のままです。主要 3クラウドのAIワークロードのうち 87〜98%はカスタマー管理の暗号化を欠いています。このギャップは測定可能で、データ上は拡大しており、介入なしに自然に閉じることはありません。

さらに、AIは運用の層に深く組み込まれています。 AI導入企業の56%が本番でエージェントフレームワークを稼働させ、64%が企業データに接続されたベクターデータベースを導入しました。そして脅威側もそこに追いついています。サプライチェーン攻撃は　AI固有のパッケージやツールを標的とし、エージェント型システムは安全制御が破られた際に現実の被害を引き起こし、研究者は武器化可能な　AI環境、プロンプトインジェクションによる水平移動、汚染された　RAGパイプラインを実証してきました。本番環境の規模は、攻撃者にとって経済的に魅力的と言えるだけの大きさに達したのです。

最後に、対応はすでに遅れています。コロラド州の改正　AI法（S.B. 26-189）は2027年1月1日に施行され、施行規則はまだ策定中です。 EU AI Actの高リスクシステムに関する義務は 2026年8月2日に発効します。本レポートで記録した設定不備や技術的な指摘は、この段階ではコンプライアンス上の責任として捉えるべきであり、EUリージョンでAIを稼働させている組織は、その解消に急いで着手すべきです。

AIはもはやクラウドインフラの未来ではありません。現在そのものです。その安全確保を、後回しにすることはできません。



Orca Securityについて

Orca Cloud Security Platformは、つくる企業のためのプラットフォームです。リスクはチームの行く先すべて、あらゆるクラウド、あらゆるサービス、そして次にリリースするあらゆるAIエージェントに付いて回ります。Orcaは深く、正確で、実行につながるコンテキストを提供し、チームが「何が重要か」を把握して迅速に動けるようにします。

Temasek、CapitalG、ICONIQ Capital、Redpoint Venturesなどの支援を受けるOrcaは、SAP、Gannett、Autodesk、Lemonade、Digital Turbineを含む数百の組織に信頼されています。

詳しくは、Orcaプラットフォームの個別デモをご予約ください

AIはもはや、呼び出すサービスではありません。AIは開発環境の全体で動き、クラウド権限を持つ自律型エージェントとして、チームに代わって行動します。セキュリティを正しく実践できる組織とは、クラウドの他の部分と同じ厳格さを自社のAI領域にも適用する組織です。セキュリティは、守る対象であるつくる人たちの速度で動かなければなりません。」

GIL GERON | ORCA SECURITY CEO兼共同創業者



©2026 ORCA SECURITY. ALL RIGHTS RESERVED.