AI Agentを活用する攻撃者と
いま組織が直面するIdentityの危機

CIsco Talosが明かす最新攻撃手法のリアルと、今すぐ実践できるID防衛策

CxO Advisor, Office of the CTO, APJC
寺下 健一

セキュリティ事業 Identity Security担当
平岡 龍弘

2026年10月27日

現実世界での自律型エージェント攻撃

Target: Hugging Face

Method: エージェントによる論理的プロービング

Target: DSEWiki

Method: 制限のバイパスと乗っ取り

Target: RubyGems

Method: 高ボリュームな登録とパッケージスタッフィング

「もはや想定の段階ではない」



AIエージェントが自ら論理を用いて制限を回避し、目的達成まで執拗に欺瞞と探索を繰り返す時代はすでに到来している。防御側の課題は、「来るかどうか」ではなく**「いかにスタックを堅牢化するか」**にシフトしている。

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco
Cisco Public

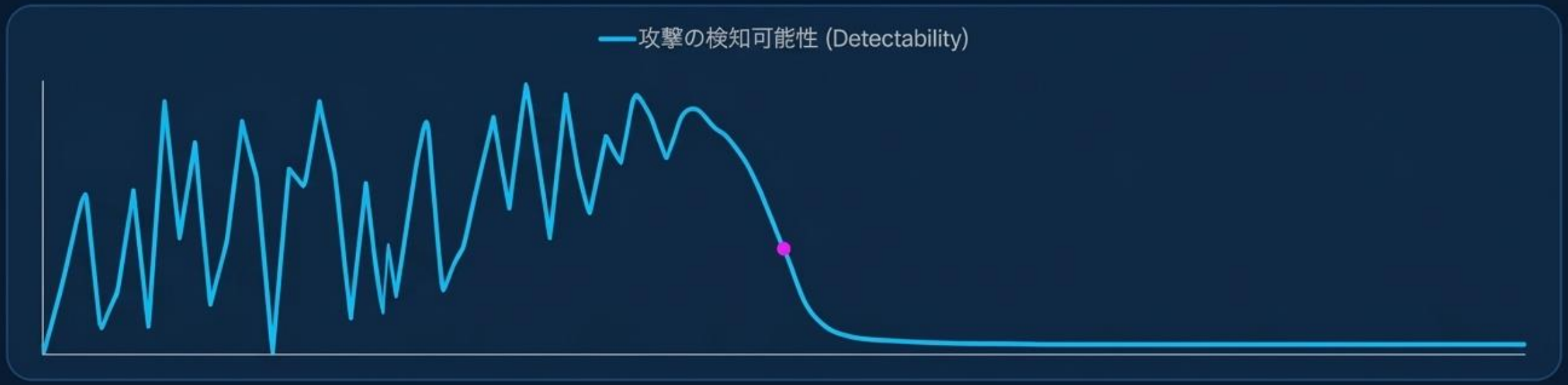
攻撃メカニズムのパラダイムシフト: Human vs. AI Swarm

	従来の人力レッドチーム (Traditional Red Team)	AIエージェントスウォーム (AI Agent Swarm)
Time (時間)	準備・実行に「数ヶ月」を要する	エージェント間のリアルタイム連携により「数時間」に短縮
Endurance (持久力)	疲労、集中力の低下、週末の休息が必要	疲労ゼロ。24時間365日、プロンプトボムを無限に継続
Preparation (事前準備)	ツールキットと個人のスキルに依存	agents.md、マークアップ指示書、ツールマッピングをロードした状態で並列実行
Scale (規模)	人的リソースによる制限	何千ものエージェントが同時に偽のアイデンティティを作成し、脆弱性を突く

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco
Cisco Public

脅威の進化プロセス：ノイズからステルスへ



現在（大音量・高ボリューム）

- ペネトレーションテストに類似。既存ツールの流用、RubyGemsのような大量スパム攻撃。
- 初期指標 (Indicators): SQLインジェクションの急増、Python / curl / wgetによるユーザーエージェント、WAF アラートのスパイク。

近未来（ステルス・低シグナル）

- 速度より「隠蔽」を優先するよう学習。永続性を確保し、機械的な持久力を持つ真のレッドチーム化。
- 脅威インサイト: ボリュームは現在のAIの特性であり、自然法則ではない。AIが静かになる前に、現在のノイズの中で検知体制を構築する必要がある。

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco
Cisco Public

AI脅威に対抗するレジリエンス・ブループリント

1

プロセスと準備 (Process & Preparedness)

インシデント対応計画（IRP）の再定義と、AI特有のシナリオを用いた机上訓練の実施。

2

アーキテクチャの堅牢化 (Architecture & Hardening)

Assumed Breach（侵害前提）のアプローチに基づく設計。エンドツーエンドでのMFAおよびFIDO2の徹底。

3

内部監視と検知 (Detection & Visibility)

境界監視からEast-Westトラフィックへの焦点移行。ネットワーク内のAIアセットのインベントリ化。

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco
Cisco Public

1. プロセスと準備：インシデント対応の再定義

IRPの運用要件



引き出しの中の計画は機能しない。明確な責任者、アウトオブバンド（帯域外）の通信手段、インシデントライフサイクル（準備～事後レビュー）へのマッピングが必須。

AI特化型テーブルトップ（机上訓練）

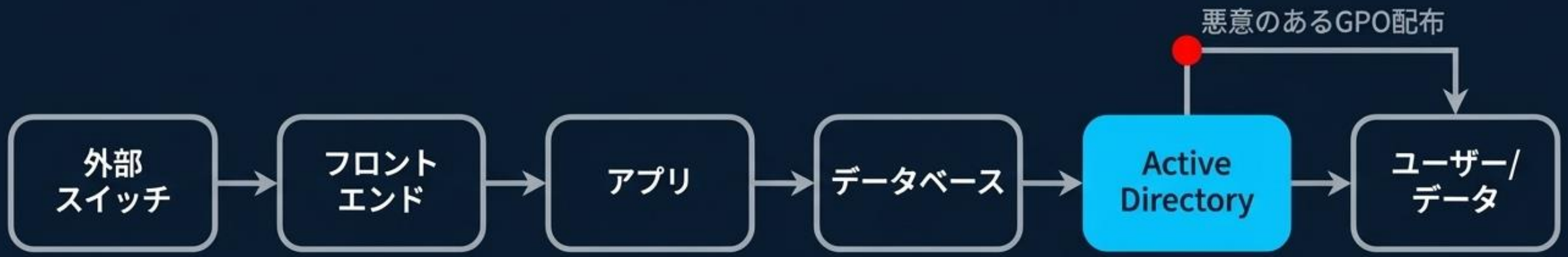
汎用的なランサムウェア訓練からの脱却。
以下のAI特化シナリオをテストする：

1. ネットワーク内をワームのように横移動し、認証情報を収集するローグAIスウォーム。
2. 自社のAIモデルの重み（Weights）データ窃取時の対応。
3. 従業員に対する大規模かつ同時のメール/SNSスプーフィング攻撃。

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco
Cisco Public

2. アーキテクチャの堅牢化：Assumed Breach（侵害前提）の設計



フットプリントの把握

ペネトレーションテストではなく、「既に足場を確保されている」前提で内部からの経路をマッピングする。

エンドツーエンドのMFA

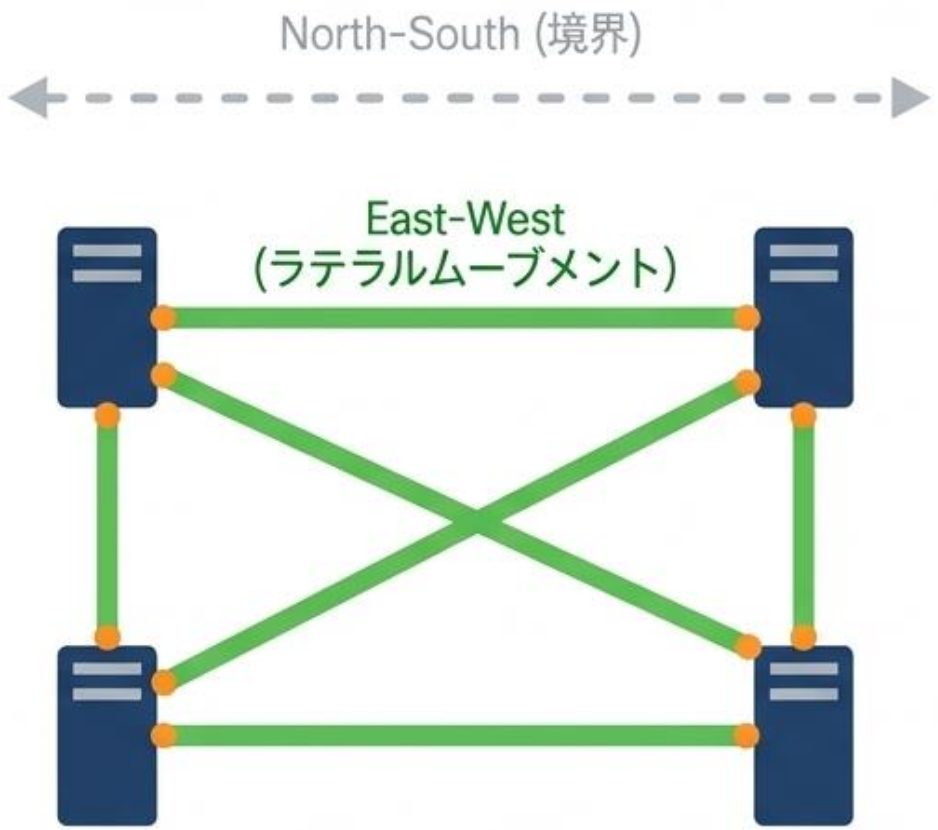
VPNのMFAだけでは不十分。ADアクセス、SSO、Linuxフリートのすべてに保護を適用。

プロンプト爆撃への耐性

執拗なAIによるSMS/Push認証の疲労攻撃を防ぐため、フィッシング耐性のあるFIDO2（セキュリティキー）を優先。

© 2026 Cisco and/or its affiliates. All rights reserved.
Cisco Public

3. 内部監視と検知：East-Westトラフィックへの焦点移行



内部インストルメンテーション: EDRの全面展開。境界（North-South）だけでなく、横移動（East-West）の可視化。C2通信やビーコニングに悪用されやすいDNSの徹底監視。

AIアセットのインベントリ管理: ネットワーク内でサーバーアクセス権を持つすべてのAIアプリケーションを監視。正規ツールであっても、コーディングAIがハイジャックされるリスクを認識する。

初期兆候の捕捉: 現在の「ノイジー」なAI攻撃フェーズ（自動化トラフィック、SQLiの急増など）を最前線で確実に検知する。

© 2026 Cisco and/or its affiliates. All rights reserved.
cisco
Cisco Public

防衛の経済学：攻撃のROI（投資対効果）を破壊する



- AIはプロンプトで指示されない限り、決して「諦める」ことはない。
- 現実的な目標は「絶対に破られないシステム（不可能）」を作ることではない。
- 攻撃の各ステップにおいて、時間、トークン、計算リソース、コストを強制的に消費させるアーキテクチャを構築する。コストを数学的に引上げることで、攻撃キャンペーンの成功確率を劇的に低下させる。

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco
Cisco Public

結論：防衛の基盤は変わらない

「攻撃の速度、持久力、そしてコスト構造は根本的に変化した。
しかし、防衛の基本原則は変わっていない。」

一部を選択的に守るのではなく、セキュリティの基本（Blocking and Tackling）を環境全体で徹底的に実行することこそが、執拗なAIスウォームの侵入を防ぐ唯一にして最大の盾となる。

インシデント対応の強化と最新の脅威インテリジェンスについては、Cisco Talosにご相談ください。

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Public


CISCO