SECURIY DOCS — No.08 トレンド

攻撃者が AI を使うと何が変わるか

AI攻撃の現在地 2026

Anthropic・Google Cloud・GMO Flatt Security の
公表資料から、守る側に求められることを整理



Securiy

12 PAGES · 2026.10
TEMPURA TECHNOLOGIES

この資料でわかること

「AI で攻撃が変わった」という話は多いものの、自社に何が関係するかは分かりにくいままです。この資料は、3社が公表した4つの資料が「実際に見たもの」を攻撃の工程で並べ直し、原文が書いている限界を拾い、自社で確かめる8項目と順番に落とします。原文は英語3本・日本語1本。課題編（No.01）P.03・P.04 を読んだ方は、P.06 から読めば重なりを飛ばせます。

想定読者

AI 攻撃が自社に関係するか判断したい開発責任者・CTO

SaaS・Fintech・EC を運営し、取引先や経営から「AI 攻撃への備えは」と聞かれる立場の方。

扱うこと

公表資料が見たものと、守る側に求められること

AI 事業者と国内の診断事業者が公表した観測。攻撃の工程ごとに AI が何をしているか、まだ変わっていないこと、自社で確かめる順番と頻度。数字は出典に譲ります。

扱わないこと

手口の再現、攻撃グループや製品の特定

攻撃の手順は書きません。特定の攻撃グループ・AI 製品・防御製品の評価や比較もしません。

1 公表資料が見たもの

3社の4資料を並べる P.03
工程ごとに、AI は何をしているか P.04
変わったこと3つ P.05
まだ変わっていないこと P.06

2 守る側に求められること

公開から悪用までの空白が縮む P.07
業務ロジックと認可の不備も、AI で見つかる P.08
新しい標的: AI 資産と開発の工程 P.09

3 自社でどう備えるか

守る側も AI を使う: 渡せる形を用意する P.10
点検表: 順番と頻度 P.11

3社の4資料を並べる: 立場の違う3社が、同じ方向を指している

AI 事業者2社と、国内の診断事業者1社。見ている場所は違いますが、「AI が攻撃の速さと規模を押し上げている」という大きな方向は4つの資料で一致しています。細かな点では、互いに反対方向の記述もあります（P.06）。

2025.11

Anthropic（AI 事業者）

商用の AI ツールを自動化の枠組みに使い、偵察から持ち出しまでを AI が実行し、人は節目の判断だけだった諜報活動を検知し公表

2026.09

Google Cloud 脅威分析チーム

先進的な攻撃者の一部が「プロンプトで聞く」段階からエージェントに工程を任せる段階へ移り、人の判断を待つ時間が縮んだと報告

2026.10

Google Cloud 脅威分析チーム

脆弱性の公開件数と悪用件数が増加。増加分の主因は公開済みの脆弱性の素早い武器化で、AI がそれを速めている可能性を指摘

2026.10

GMO Flatt Security（国内の診断事業者）

AI 診断で見つかった脆弱性の多くが業務ロジック・認証認可の不備。公開モデルが特定の課題で最上位モデルに迫ったと報告

出典の読み方

3つは各社が自社のサービス・顧客・調査で見たもの。Google Cloud 2026.10 は公開された脆弱性全体の集計に、自社の悪用観測を重ねたものです。

AI 事業者は守る側の AI 利用を、国内の診断事業者は AI 診断を提供する立場で、どちらも自社の製品に近い結論になりやすい資料です。
この資料を出している Securiy も AI を使った検証を提供しているので、原文が書いている限界（P.06）も同じ重さで並べます。

件数・割合・所要時間は原文にあります。本文では出さず、裏表紙に資料名を載せます。

4つの資料から読み取れる変化を、この資料では「速く・何度も・安く」の3語で整理します。各点の出典は P.05 に記します。

工程ごとに、AI は何をしているか

2社の資料に書かれている観測を、攻撃の工程ごとに並べました。● はその資料に観測の記述があるもの、— は記述がないものです（観測されなかった、という意味ではありません）。

攻撃の工程	Anthropic 2025.11	Google Cloud 2026.09	公表資料に書かれていること（要約）
偵察・標的の選定	●	●	標的を選ぶのは人。どこが弱いかを調べるのは AI
脆弱性の調査・攻撃コードの作成	●	●	AI が脆弱性を調べ、攻撃コードを書く。公開済みの脆弱性を使える形にする速度も上がっている
侵入・認証情報の収集	●	●	エージェントがスキャンと収集をつなげ、途中のエラーも自分で直す
横展開・データの持ち出し	●	●	横展開（入った先から別のシステムへ広げる）の手順の作成と、持ち出したデータの要約に AI を使う
マルウェアの作成・難読化	—	●	検知を避けるための書き換えに AI を使う
文面の作成（だまし）	—	●	標的に合わせた誘い文を AI が作る
攻撃の記録・次の段階の計画	●	●	攻撃の記録や持ち出したデータの要約を AI が書き、次の段階の計画に使う
AI 事業者の安全策の回避	●	●	目的を伏せて小さな作業に分ける。盗んだアカウントや公開モデルを使う

読み方

最も自動化が進んだ事例では、人が残っていたのは「標的を選ぶ」「続けるかを決める」の節目でした。多くの攻撃グループは、人が AI に作業を頼む使い方にとどまる、と同じ資料は書いています。

自社に当てはめると

「うちは標的に選ばれない」は、標的を選んでから始まる攻撃にだけ当てはまる話です。公開済みの弱点を使う攻撃は、該当する構成を広く探すところから始まります（P.07）。

変わったこと3つ: 何度も、速く、安く

4つの資料から読み取れる3点です。各点を支える出典は1〜2本で、出典の中に反対方向の記述がある点は P.06 に書きます。

1
何度も: 人の判断を待たない

以前

手順の一つひとつを人が操作
途中で止まり、人の判断を待つ

いま（公表資料）

AI が工程をつなげて実行
最も進んだ事例では、人は節目の判断だけ

Anthropic 2025.11
Google Cloud 2026.09（先進的な攻撃者の一部）

2
速く: 公開から悪用まで

以前

公開された弱点の解析に日数
攻撃コードを人が書く

いま（公表資料）

公開から数日で複数グループが悪用した例
AI による差分解析が速めている可能性

Google Cloud 2026.10（1件の事例）
GMO Flatt Security 2026.10（差分解析の実験）

3
安く: 手元のモデルで足りる

以前

高性能な AI は事業者の API 経由
利用規約と安全策の内側

いま（公表資料）

公開モデルが特定の課題で最上位に迫った
安全策の外で動かせ、攻撃側も同じ条件

GMO Flatt Security 2026.10
ただし費用はなお障壁（Google Cloud 2026.09）

どれも新しい手口ではありません。これまでの手口が、速く・何度も・安く試されるようになった、という変化です。

まだ変わっていないこと: 資料が書いている限界

同じ資料は、変わっていないことも書いています。4点とも原文に記述があります。守る側に加わるのは、新しい種類の弱点ではなく、
確かめる順番と頻度の見直しと、AI を使う会社が持つ資産（P.09）です。

1 AI は、でっち上げる

攻撃に使われた AI は、存在しない認証情報を作ったり、公開されている情報を
「秘密を抜き出した」と報告したりしました。AI 事業者自身が、これを
完全に自律した攻撃の障害として挙げています。

Anthropic 2025.11

2 完全な自律は、まだ観測されていない

公表された事例でも、標的の選定と節目の判断は人が行いました。脅威分析チームは
「完全に自律した攻撃の流れが実際の標的に使われた例はまだ観測していない」と書き、
段階的な成熟と表現しています。ただし同じ資料は、その方向へ進みつつあるとも
書いています。

Anthropic 2025.11 ／ Google Cloud 2026.09

3 悪用の増加分は、ゼロデイではなく公開済みの弱点から

悪用の件数ではゼロデイ（修正版が出る前に悪用される脆弱性）がなお多数です。
増加分の主因は公開済みの脆弱性の素早い武器化で、AI がそれを速めている可能性が
指摘されています。公開された脆弱性のうち実際に悪用されるものはごく一部で、
優先順位づけが要ります。

Google Cloud 2026.10

4 AI 事業者の監視に見つかり、止められている

悪用は AI 事業者の監視で検知され、アカウントは停止されました。ただし Anthropic の
事例では、安全策は作業の分割と偽装で回避され、一部の侵入は成功していました。
高価なモデルの利用料と計算資源はなお障壁で、だから AI アカウントの窃取や
クラウドの乗っ取りが増えている、とも書かれています。

Anthropic 2025.11 ／ Google Cloud 2026.09

「AI で何でも破られる」でも「AI の攻撃はまだ使い物にならない」でもありません。既知の脆弱性の放置と業務操作の抜け、という従来の弱点に、速さと回数が加わりました。

公開から悪用までの空白が縮む: 照合を「月1」から「公開のたび」に

攻撃側が公開済みの弱点を日単位で使う例があるなら、守る側の「公開を知る → 該当を決める → 修正版を当てる」も、月1回の棚卸しでは間に合いません。該当したものは悪用の観測を待たずに被害を絞って待ち、修正版は悪用が観測されたものから当てる。これが公表資料の示す順番です。



自分でやるなら

公開を知る経路と該当を決める人を先に決め、構成表（SBOM: 使っている部品と版の一覧）を持つと、月1回の棚卸しを待たずに公開のたびに照合できます（No.10）。

修正版は、悪用が観測されたもの（KEV: 米 CISA が公開する悪用確認済みの一覧。載っていない＝悪用されていない、ではない）と、外部に公開された機器・認証のない管理画面から。

当てるまでの間は、該当機能の停止・外部からの接続制限で被害を絞り、アプリの操作を経由する被害なら上限も下げて待つ（No.19 P.06・P.07）。

業務ロジックと認可の不備も、AI で見つかるようになった

「他組織のデータが見えないか」「承認を迂回できないか」は、スキャナでは見つからず、人の診断の範囲でしか試されてきませんでした。守る側の AI 診断がこの種類を多く見つけたという公表は、攻撃側の AI も今後、同じものを見つけうる、という意味です。

従来型のスキャナが見るもの

- 既知の脆弱性の型に当てはまる入力（注入・スクリプト混入）
- 公開ページの設定（応答ヘッダー・証明書・古い版）
- ライブラリの既知の脆弱性との照合（別のツールで行う）
- 単独の要求で判定できるもの

AI 診断が多く見つけた種類: 業務ロジック・認証認可の不備

（GMO Flatt Security 2026.10。下の4つは説明のための例）

- 他組織の ID を指定したデータの取得
- 承認を飛ばした実行、承認内容の書き換え
- 上限を超える出力・送金、連携キーの範囲外
- 複数の要求を組み合わせないと分からないもの

何が変わったか

業務ロジック・認可の不備は、従来型のスキャナでは検知が難しく、人の診断でしか見つからない種類でした。
国内の診断事業者は「まだ十分な診断を受けていないシステムにおいては、このような種別の脆弱性が、今後 AI により自律的に発見・悪用されるリスクがある」と書いています。
攻撃側も同じ種類の AI を使えるなら、「診断済み」の範囲の外（見積りに入れなかった経路、診断後に足した操作、連携キー・AI 機能の経路）が、AI に試される側に回ります。

自分でやるなら: 検証環境に2つの組織を作り、一方の ID で他方のデータを画面・API・CSV から取りに行く手順は No.09 P.08・P.09 にあります。

新しい標的: AI 資産と、開発の工程

攻撃側が AI を使うには AI へのアクセスと計算資源が要るため、AI のアカウント・キー・計算資源と、AI を使った開発の工程が標的になっています。自社で AI を使っていれば、守る対象がここにも増えています。

標的	公表資料が見たもの（Google Cloud。括弧内は資料の年月）	自社で確かめること
AI の API キー・アカウント	開発者の認証情報の窃取、侵害済みアカウントの売買、企業のクラウドを乗っ取って AI の計算資源を使う（2026.09）	**キーの範囲・期限・利用量の上限と通知。使っていないキーの失効（No.11）**
自社のモデル・プロンプト・学習データ	AI 企業に限らず、医療・行政・メディアなど AI を使う組織も標的に。諜報だけでなく、盗んで脅す目的でも（2026.09）	**どこに置いてあり、誰が読めるかの一覧。保管先の権限と記録**
AI コーディング支援・開発の工程	正規の署名つきで汚染版パッケージを公開し、AI の自動確認を通す。AI が悪性の依存を取り込んだ例、設定ファイルの指示で AI にコマンドを実行させる例（2026.09）	**依存の追加はロックファイルとレビューを通す。AI エージェントの実行環境を隔離する**
自社の AI 機能が使う基盤	エージェント基盤・AI ゲートウェイ・推論基盤の脆弱性公開が急増し、公開済みの AI ミドルウェアの脆弱性が実際に悪用された（2026.10）	**AI 機能が使う基盤の版と公開情報の照合（No.10）。AI 機能からの要求を「誰として」照合するか（No.11 P.07）**

用語

ロックファイル: 依存パッケージの版を固定する記録。AI エージェントの隔離: AI が実行するコマンドや通信を、本番の認証情報・秘密情報・許可していない外部通信に届かない環境に閉じ込めること。

読み方

4行とも「AI の攻撃」ではなく「AI を使う会社が持つ資産」の話です。右列の多くは、クラウドや AI サービスの設定から始められます。

守る側も AI を使う: 製品より先に、渡せる形を用意する

3社とも、守る側が AI を使うことを勧めています。国内の診断事業者は、検証対象・実行環境・予算の三つを整える必要があり、「自社の情報を AI に渡せる利点は、渡せる形を自社で用意して初めて生かせます」と書いています。

三つの調達

脆弱性情報
自社のどこに弱点があるか

脅威情報
自社に届きうる攻撃は何か

AI の推論能力
商用でも公開モデルでも

三つの準備（いずれも GMO Flatt Security 2026.10、補足は要約）

検証対象
試す操作と環境の一覧

実行環境
AI が安全に試せる場所

予算
変更のたびに回す前提で

→

都度 AI に任せる → 開発の工程に組み込み、変更のたびに

AI 事業者も、守る側が AI を試す領域として、監視の自動化・脅威検知・脆弱性評価・インシデント対応を挙げています（Anthropic 2025.11）。

脅威分析チームは、脆弱性の優先順位づけと修正に AI 支援の仕組みを使うこと、開発元が出荷前に AI 支援のコードレビューを回すことを勧めています（同 2026.10）。

どの AI 製品を使うかは、この資料では扱いません。
製品より先に、渡せる形を用意するのが順番です。

自分でやるなら:「検証対象」を、渡せる形3つにする

- **① 検証環境とテストアカウント**
- **② 守りたいデータに触れる操作の一覧**
- **③ 使っているライブラリと版の一覧（構成表）**

この3つがあれば、商用の AI 診断でも、手元で動かす公開モデルでも、外部の検証でも、同じ入力で回せます。予算は、変更のたびに回す前提で組みます。

点検表: 速く・何度も・安く試される前提で、8つを順番に

公表資料が示す変化に対して、自社で確かめる8項目です。迷ったら、Q1・Q2（公開を知り、照合する）→ Q3（待つ間の手順）→ Q4（画面以外の経路）→ Q5（AI のキー）→ Q6〜Q8 の順で。頻度の目安は「目安と頻度」の列にあります。

No.	点検すること	目安と頻度	状態	担当
Q1	**公開された脆弱性を知る経路と、該当を決める人が決まっている**	ベンダー通知・JVN・KEV を誰が見るか（No.10）。公開のたび	ある ／ 一部 ／ ない	
Q2	**構成表（SBOM）があり、公開のたびに照合できる**	ロックファイルから自動で作れる。遅くとも1週間以内	ある ／ 一部 ／ ない	
Q3	**修正版を当てるまでの間に、該当機能を止める手順と、上限を下げる手順がある**	誰が・どこから・どう戻すか（No.19）。平時に一度試す	ある ／ 一部 ／ ない	
Q4	**他組織の ID・承認の迂回・上限の超過を、画面以外の経路でも試したことがある**	API・CSV・連携キー・AI 機能（No.09）。 操作を足したリリースごと	ある ／ 一部 ／ ない	
Q5	**AI の API キーに範囲・期限・利用量の上限と通知がある**	使っていないキーは失効している。四半期ごと	ある ／ 一部 ／ ない	
Q6	**自社のモデル・プロンプト・学習データの置き場所と、読める人が分かる**	保管先の権限と記録。四半期ごと	ある ／ 一部 ／ ない	
Q7	**依存の追加はレビューを通り、AI エージェントの実行環境が隔離されている**	本番の認証情報・秘密情報に届かない。四半期ごと	ある ／ 一部 ／ ない	
Q8	**AI に渡せる形（検証環境・操作の一覧・構成表）がある**	P.10 の3つ。年1回見直す	ある ／ 一部 ／ ない	

自社で回るもの

Q1〜Q3・Q5〜Q8 は、この資料と No.10・No.19・No.11 の手順で自社でできます。Q4 も、決めた操作を画面・API・CSV から試すところまでは自社で（P.08）。外部に出す前に、年次診断の範囲に業務操作の経路が入っているかを見積書で確かめてください（No.09 P.06、確かめる5項目は No.26 P.02）。入っていれば、その診断で足ります。

外に出す価値があるもの

Q4 の、まだ決めていない条件と経路（承認内容の書き換え・連携キー・AI 機能）を、攻撃側の AI と同じように広く洗い出して試し、試さなかったものも未検証範囲として書いた記録にするところ。共同検証プログラムでは、守りたいデータに触れる操作から2操作を選びます。

速く、何度も、安く試される。
だから、確かめる順番を決める。

公開された脆弱性の照合、待つ間の手順、AI 資産の棚卸し、渡せる形の準備と、業務操作を決めた条件で画面と API から試すところまでは自社でできます（No.09 P.09）。外に出す価値があるのは、まだ決めていない条件と経路（承認内容の書き換え・連携キー・AI 機能）を広く洗い出して試し、試さなかったものも未検証範囲として書いた、取引先に出せる記録にするところです。初回のご相談は30分・無料です。

お申し込み・お問い合わせ　**securiy-preview.pages.dev**　のフォームから
Tempura technologies株式会社

※ 本資料は公表資料の要約で、各社の見解を代弁するものではありません。すべての攻撃の網羅や「安全」の総合判定は保証しません。金融の規制適合審査、決済基盤全体の監査、カード不正判定は対象外です。

無料相談

30分無料相談

点検表の Q4 を一緒に埋め、守りたいデータに触れる操作から初回に検証する2操作を決めます。本番の認証情報や顧客データは不要です。

相談を申し込む →

共同検証プログラムとは

検証テーマを一緒に作る企業に、初回スコープで提供する枠組み

1アプリ・1検証環境・2操作、準備完了後10営業日・合計40時間（修正案と再検証は2件・12時間まで）。費用は着手前に提示します。

出典

本文で参照した公表資料

Anthropic（2025.11）Disrupting the first reported AI-orchestrated cyber espionage campaign

Google Cloud GTIG（2026.09）AI Threat Tracker: From Prompting to Autonomy

Google Cloud GTIG（2026.10）Vulnerability Discovery and Exploitation Trends in the AI Era

GMO Flatt Security（2026.10）AIによるセキュリティ診断・ペネトレーションテストの現在地 v2026.10