Monthly Topics 2026.09

・AI 関連：攻撃者による AI エージェント悪用の本格化と防御側の動き

・攻撃手法：ClickFix が OS からブラウザへ、暗号資産を狙う

① AI 関連

Claude AI エージェントの悪用：標的は拡大　/ オペレーターは削減

出典：Detecting and countering misuse of AI: September 2026 - Anthropic (2025.12〜2026.8)

GTG-20006

ロシア政府との関連が疑われる

- 標的：ウクライナ・欧州の政府機関、防衛企業、軍用ドローンのサプライチェーン
- インフラ取得・フィッシング用ドメイン登録・C2 監視をエージェントで自動化
- ドローン部品メーカーのメールボックス一括エクスポート、SDK 窃取

北アフリカ：身分証明記録　30万件超／登記データ　50万社超の窃取疑い

GTG-10007

中国語話者のオペレーター

- 脆弱性調査とエクスプロイト開発に特化
- AI エージェントで脆弱性の発見からエクスプロイト生成までを反復

1ヶ月で 12件超のゼロデイ・エクスプロイトを生成

ShinyHunters 関連

金銭目的

- Android アプリ・コードリポジトリ・クラウド・企業システムから認証情報を収集
- APK を逆コンパイルし、ハードコードされたシークレットを探索

EC2 10台で Android APK 180万個を解析

共通点：手法は従来どおり、実行はマシンスピードへ　→ 標的の範囲は拡大し、必要なオペレーターは削減

① AI 関連

エージェンティック AI の武器化：偵察・悪用・侵害後を自動化

出典：Hackers Weaponize Agentic AI to Automate Reconnaissance, Exploitation and Post-Exploitation - gbhackers

攻撃者の構成　LLM ＋ スキャナ／資産探索エンジン／コードリポジトリ／プロキシ／エクスプロイトFW／シェル実行

1 偵察
OSINT 収集、ドメイン・クラウド資産のマッピング、CVE 突合と優先順位付け

→

2 悪用
脆弱性評価・エクスプロイト開発、認証の突破

→

3 認証情報の窃取
クラウドトークン・API キー・セッション Cookie の収集

→

4 侵害後
内部資産の列挙、横展開の計画、ペイロード準備

< 6時間
大規模な認証情報収集基盤を構築・実行（Mandiant, 2026 Q2）

23,800+
公開 C2 で管理されていた収集済みシークレット（クラウド／AI の API キー等）

事例 - knaithe（中国語話者）

DeepSeek＋Hermes Agentで偵察・脆弱性探索を自動化。
独自の「1DayNews」でRCE情報を収集・評価し標的候補を通知。
Langflow → n8nへ自律的にPivotしたが認証等でExploitは阻害され、
その後オペレーターが Citrix NetScaler／Marimo等への手動攻撃へ移行。
CitrixではSession Cookie窃取によるMFA回避、
MarimoではAWS認証情報窃取とNKAbuse投入を確認。

現状：完全自律ではなく「人間監督下の半自律型」 — それでも人手の確認サイクルでは追いつけない

① AI 関連

OpenAI Astra：サイバー能力レベル Critical に到達した初のモデル

出典：OpenAI Reveals Astra, Its First AI Model to Reach 'Critical' Cybersecurity Risk Threshold - Security Boulevard (2026.9.2)

Astra の概要

位置付け	OpenAI 内部のリスク評価で、最高の脅威レベル Critical に初めて到達
Critical の目安	適切なツールとアクセスがあれば、人の逐次介入なしに未知の脆弱性を発見し、その悪用手法を開発できる
経緯	Critical到達を踏まえ、開発・リリースの一部を遅延→監視・隔離・アラインメント等を強化して段階的に再開
安全対策	Chain-of-Thought監視で未承認・逸脱行動を検知し、自動停止

専門家の指摘（要旨）

安全対策の強化は前向きだが、アクセスを守ることは責任の一部にすぎない

— Stephanie Walter（HyperFRAME Research）

評価で示された攻撃能力

100%

ExploitBench のスコア

2件

High 脆弱性 20件のテストで特定・悪用したゼロデイ

→ ブラウザ侵害チェーンの完全構築とサンドボックス・エスケープ、一般ユーザー権限から root への昇格も実証

提供方針：高度な機能は段階的に展開

1	**選抜テスター**	高度なサイバー機能をまず限定提供
2	**Daybreak Blue**	OpenAI のサイバーセキュリティ向け早期アクセス
3	**一般公開版**	近日公開予定（制限付き）

サイバー脱獄プロンプトの拒否率 **59%** → **91.5%**（前世代比）

① AI 関連

OWASP OASIS：AI × 専門家で OSS の脆弱性を迅速に FIX

出典：gbhackers｜IoT OT Security News (2026/09/01) ※ 発表 2026/08/26

課題：脆弱性の発見スピードに OSS メンテナーの修正が追いつかず、攻撃者は AI でその隙を突く

1 検出・生成

AI

- 広く使われるリポジトリを AI がスキャン
- セキュリティ上の問題を特定し、修正パッチ候補を自動生成

発見と修正案をセットで

→

2 検証

AppSec 専門家＋自動エージェント

- 専門家・エージェント・コミュニティがパッチの安全性・正確性を検証

数日〜数週間 → 数分

→

3 共有

アップストリームのメンテナー

- 検証済みの修正を提供し、機能・性能をテスト
- 採用・修正・却下の最終判断はメンテナーが保持

人間の判断を残す

創設メンバー： AppSecAI／Intigriti／DryRun Security ＋ 複数業界の数百人の AppSec 専門家が参加

位置付け：Patch the Planet（OpenAI）・Akrites（LF）・Project Glasswing（Anthropic）と補完し、幅広い依存 OSS をカバー

① AI 関連

攻撃は「マシンスピード」へ：防御側に求められること

今月の AI 関連 4 記事に共通するポイント

⚔️ 攻撃側で起きていること

1. **偵察〜侵害後の自動化**
   少ない人員で多数の標的へ（GTG-20006 ほか）
2. **ゼロデイ発見・悪用の高速化**
   GTG-10007／Astra の評価結果
3. **AI の API キー・トークンも標的に**
   収集済みシークレット 23,800件超
4. **開発・AI 基盤ツールが侵入口に**
   Langflow／n8n／Marimo、Citrix

🛡️ 防御側でやるべきこと

1. **AI API キー・トークンを本番シークレット同等に管理**
   最小権限・漏洩時のローテーション・異常利用の監視
2. **外部公開の管理 IF を Zero Trust の背後へ**
   インターネットから直接触れさせない
3. **n8n／Langflow／Notebook 環境を厳格に制限**
   処理環境の隔離もあわせて
4. **「スキャン→取得→悪用」の高速な連鎖を検知**
   人手の確認を待たず自動で封じ込め

今回の構図：攻撃にも防御にも AI エージェントが入り、「速度」の競争に　— OASIS や Daybreak Blue など防御側の活用も加速

② 攻撃手法

ClickFix の変化：OS からブラウザへ、狙いは暗号資産

出典：Cisco Talos (2026/09/08)／Infosecurity Magazine｜IoT OT Security News (2026/09/09)

ClickFix は、偽のエラー表示などでユーザーをだまし、ユーザー自身にコマンドを貼り付けて実行させる手口。

今回 Cisco Talos が報告したキャンペーンでは、実行の場が　OS からブラウザへ移った。被害者に　Tampermonkey 拡張機能とスクリプトを導入させ、公開 Google スプレッドシートから難読化コードを取得。暗号資産取引サイトのセッション内に注入し、資金を盗み出す。

キャンペーンのタイムライン

- **2025/10** 開始：Chrome のアドレスバーに JavaScript を貼り付けさせる
- **2026/03** Google Visualization API で公開スプレッドシートからコード取得
- **2026/04** Tampermonkey の導入を指示 Talos 通知 → 1週間後に新シートで再開
- **2026/08** Google ドキュメントも悪用し活動継続（8/11 時点）

🎣 ルアー：「存在しない脆弱性」で釣る

- Telegram／DarkForums／テキスト共有サイトに月2回以上投稿
- 暗号資産スワップサービスの「存在しないAPI 脆弱性」の漏洩レポートを装い、最大 38% の利益を謳う
- 狙いは「理解していない脆弱性でも悪用しようとする人物」

49	約0.159	3,000+
BTC アドレス	BTC（約1万ドル）	アドレス＋ミキサーで洗浄

② 防御側への示唆

ブラウザ内で完結する攻撃：拡張機能と正規サービスの通信を見る

「キャンペーン自体は大半の組織の脅威ではないが、使われている手法は脅威となる」（記事の評価）

1 ルアー投稿
偽の脆弱性レポートで誘う

→

2 自ら導入
Tampermonkey＋スクリプトを被害者が追加

→

3 コード取得
Google Sheets のセルに白文字で隠蔽
第2段 21種・XOR を定期更新

→

4 セッションへ注入
fetch API を改変し入金アドレスを差し替え

→

5 窃取
クリップボードのアドレスも置換

従来の ClickFix vs 今回のキャンペーン

比較軸	従来の ClickFix	今回のキャンペーン
実行の場所	OS（「ファイル名を指定して実行」・ターミナル）	ブラウザセッション内
コードの置き場所	攻撃者が用意したサーバー等	正規サービス（Google Sheets／Docs）
見るべきポイント	プロセス起動・コマンド実行	拡張機能・ブラウザからの通信

→ OS 側のプロセス監視だけでは捉えにくい

✅ 必要になる対策

- 役割に応じてブラウザ拡張機能を制限（許可リスト化）
- ブラウザセッション内からの想定外のGoogle Docs／Sheets へのリクエストを監視
- 「コードを貼り付けて実行」「拡張機能を入れて」という指示そのものを疑うよう利用者へ周知

手法そのものは他の目的にも転用可能　— 業務端末でも同様の監視を

参照記事一覧

1. Claude AI エージェントを悪用する攻撃者：標的の拡大とオペレーターの削減を達成
   iototsecnews.jp 2026/09/12 （原文：gbhackers.com／Anthropic 脅威インテリジェンスレポート）

2. エージェンティック AI の武器化：偵察行動／エクスプロイト／侵害後の自動化が進行している
   iototsecnews.jp 2026/09/09 （原文：gbhackers.com／GTIG・Mandiant・Deepwatch）

3. OpenAI Astra が発表：脅威レベル Critical に到達する初のモデルへの期待と懸念
   iototsecnews.jp 2026/09/02 （原文：securityboulevard.com）

4. OWASP が立ち上げた OASIS：オープンソースの脆弱性を AI と専門家の知見で迅速に FIX
   iototsecnews.jp 2026/09/01 （原文：gbhackers.com）

5. ClickFix キャンペーンの変化：標的を OS からブラウザへ移行しながら暗号資産を狙う
   iototsecnews.jp 2026/09/09 （原文：infosecurity-magazine.com／Cisco Talos 2026/09/08）