AIセキュリティ短信 2026年6月号

※注意※

本稿はAIセキュリティに関連するニュースを「AIに係る安全性確報（Security for AI)」・「AIを活用したサイバーセキュリティ確保（AI for Security)」・「AIを悪用したサイバー攻撃への対処」という3つの観点で収集・選別し要約したものです。個々の事例の記載内容は参照している情報源の記載に準じたものであり、その内容の正確性・妥当性等をIPAにおいて保証するものではなく、また、参照される製品・ソリューション等を批判・推奨するものでもありません。

目次

本 AI セキュリティ短信は AI やセキュリティに関するご知見をお持ちの IPA 外の方々のご協力・ご確認を経て編纂いたしました。改めて皆様に御礼申し上げます。

1. AIに係る安全性確保（Security for AI）

1.1. [2026-03-16] AIエージェントによるスプレッドシートデータ流出脆弱

Claude for Excel や Ramp の Sheets AI などの AI 統合スプレッドシートツールにおいて、間接的プロンプトインジェクションの脆弱性が確認された。攻撃者は外部データセットに悪意ある指示を埋め込み、AI エージェントを操作して、機密データを外部サーバーに送信する数式（=IMAGE 関数等）を生成させる。これら攻撃は、正規のスプレッドシート機能を悪用し防御を回避する。Anthropic や Ramp は警告機能やユーザー確認プロセスを導入したが、依然として不十分である。専門家は「リンクされたデータ型」の制限、信頼できない外部データの利用禁止、およびユーザー教育を推奨している。

情報源

- "Ramp’s Sheets AI Exfiltrates Financials" (2026-03-16)
  https://www.promptarmor.com/resources/ramps-sheets-ai-exfiltrates-financials
- "CellShock: Claude AI is Excel-lent at Stealing Data"
  https://www.promptarmor.com/resources/cellshock-claude-ai-is-excel-lent-at-stealing-data

1.2. [2026-03-31] Axios npm パッケージ侵害によるサプライチェーン攻撃

2026年3月末、広く利用されている Axios npm パッケージ（v1.14.1 および 0.30.4）が、メンテナーアカウントの乗っ取りにより侵害された。攻撃者は悪意ある依存関係「plain-crypto-js@4.2.1」を挿入し、postinstall フックを通じてクロスプラットフォームの RAT（WAVESHAPER.V2）を展開した。国家系脅威アクターUNC1069 の関与が疑われるこの攻撃は、難読化されたドロッパーとプラットフォーム別ペイロード（PowerShell、C++、Python）を用い、システム情報の窃取や遠隔コマンド実行を行った。悪意あるバージョンは約3時間で削除されたが、マルウェアはホスト環境を完全に侵害するため、影響を受けた環境の認証情報やシークレットの全刷新が強く推奨される。Axios 自体は AI システム等にあたらないが、Claude Code などの AI 製品でも広く利用されており、AI システムのサイバー脆弱性を狙ったサプライチェーン攻撃の一種になっている。また、攻撃過程では Axios のメンテナを騙すためにディープフェ

イクなどを駆使したソーシャルエンジニアリングが用いられた形跡があり、AIを悪用したサイバー攻撃の側面も有する事例となっている。

情報源

- "Axios NPMパッケージ侵害：週1億以上のダウンロードを誇るJavaScript HTTPクライアントにサプライチェーン攻撃 | トレンドマイクロ（JP)" (2026-04-01) https://www.trendmicro.com/ja_jp/research/26/d/axios-npm-package-compromised.html
- "Inside the Axios supply chain compromise - one RAT to rule them all — Elastic Security Labs" (2026-04-01) https://www.elastic.co/security-labs/axios-one-rat-to-rule-them-all
- "North Korea-Nexus Threat Actor Compromises Widely Used Axios NPM Package in Supply Chain Attack | Google Cloud Blog" (2026-04-01) https://cloud.google.com/blog/topics/threat-intelligence/north-korea-threat-actor-targets-axios-npm-package/?hl=en
- "Post Mortem: axios npm supply chain compromise · Issue #10636 · axios/axios" (2026-03-31) https://github.com/axios/axios/issues/10636

1.3. [2026-03-31] Anthropic「Claude Code」のソースコード流出

Anthropic社は、npmパッケージ内にソースマップファイル(.map)が誤って含まれていたことにより、AIコーディング支援ツール「Claude Code」のソースコードを流出させた。流出した約51万行のTypeScriptコードはGitHub上で拡散され、これを元にAI支援によりわずか1日で開発されたクローンが登場する一方で、この騒動を悪用した攻撃者がマルウェア（Vidar情報窃取ツールやGhostSocksプロキシなど）を混入させた「リーク版ソース」を配布している。開発者は非公式リポジトリの使用を避け、ゼロトラストアーキテクチャの導入と公式の署名済みバイナリのみを利用することが強く推奨される。この事例はAIシステム開発を手掛けるビッグテックが人為的ミスにより引き起こしたセキュリティ事故である。

情報源

- "Anthropic accidentally exposes Claude Code source code" (2026-03-31) https://www.theregister.com/software/2026/03/31/anthropic-accidentally-exposes-claude-code-source-code/5227940
- "Anthropic leaks part of Claude Code's internal source code" (2026-03-31) https://www.cnbc.com/2026/03/31/anthropic-leak-claude-code-internal-source.html
- "Claude Code Leak: Critical AI Security Threat 2026" (2026-04-01) https://www.zscaler.com/blogs/security-research/anthropic-claude-code-leak
- "Claw Code vs Claude Code (2026): Open-Source GitHub Clone & Archi" (2026-04-01) https://www.eigent.ai/blog/claw-code

1.4. [2026-04-21] Anthropic「Mythos」に対する不正アクセス

Anthropic の次世代高性能 AI モデル「Mythos」が、サードパーティベンダー環境を介して不正アクセスを受けた。本モデルは高度な推論とサイバー脆弱性の特定能力を有しており、以前には CMS の設定ミスによる大規模なデータ漏洩でもその存在とリスクが露呈していた。Mythos は防御と攻撃の双方に転用可能な「ステップチェンジ」的な能力を持つため、今回の不正アクセスは深刻なセキュリティ上の懸念をもたらしている。Anthropic は現在「Project Glasswing」を通じて、防御目的での利用を前提とした限定的なリリースを制御している。

情報源

- "Anthropic’s Mythos AI Model Is Being Accessed by Unauthorized Users - Bloomberg" (2026-04-22) https://www.bloomberg.com/news/articles/2026-04-21/anthropic-s-mythos-model-is-being-accessed-by-unauthorized-users
- "Anthropic's Mythos model accessed by unauthorized users, Bloomberg News reports | Reuters" (2026-04-22)

https://www.reuters.com/technology/anthropics-mythos-model-accessed-by-unauthorized-users-bloomberg-news-reports-2026-04-21/

- "Exclusive: Anthropic 'Mythos' AI model representing 'step change' in power revealed in data leak | Fortune" (2026-03-26)
  https://fortune.com/2026/03/26/anthropic-says-testing-mythos-powerful-new-ai-model-after-data-leak-reveals-its-existence-step-change-in-capabilities/

1.5. [2026-04-09] Anthropic「Claude Cowork」の一般リリースと「Claude Code」強化

Anthropic は Claude Cowork を正式に一般リリースすると共に、Claude Code を拡張し、プラットフォーム間同期、権限管理用の「オートモード」、およびデスクトップ UI の直接操作機能を追加した。これにより、Claude はローカルアプリやファイル、ブラウザを直接操作してタスクを実行可能となる。生産性は向上する一方、不正なファイル操作やプロンプトインジェクション、機密データへの過剰なアクセスといったセキュリティ上の懸念が生じる。防御策として、厳格なロールベースアクセス制御（RBAC）、OpenTelemetry を通じた SIEM 連携による監視、アプリごとの詳細な権限管理が不可欠となる。高リスクなユースケースにおけるリスク判断の重要性も高まっている。

情報源

- "Assign tasks from anywhere in Claude Cowork | Claude Help Center" (2026-04-25)
  https://support.claude.com/en/articles/13947068-assign-tasks-from-anywhere-in-claude-cowork
- "Auto mode for Claude Code | Claude" (2026-03-24)
  https://claude.com/blog/auto-mode
- "Let Claude use your computer in Cowork | Claude Help Center" (2026-04-25)
  https://support.claude.com/en/articles/14128542-let-claude-use-your-computer-in-cowork
- "Making Claude Cowork ready for enterprise | Claude" (2026-04-09)

https://claude.com/blog/cowork-for-enterprise

1.6. [2026-04-19] Vercel への OAuth サプライチェーン攻撃

Web アプリケーションのデプロイプラットフォームとして知られる Vercel が 2026 年 3 月から 4 月にかけてサプライチェーン攻撃を受けた。この攻撃は別のサービスである Context.ai の従業員がそれと気付かずマルウェアをインストールしたことに始まる。Vercel 従業員は、Context.ai を利用するために、自身の Google Workspace に対するアクセス権を Context.ai に付与しており、これが Context.ai 側に OAuth トークンとして保存されていた。攻撃者は Context.ai を侵害したマルウェア（Lumma Stealer）によってこれらの OAuth トークンを盗み出し、Vercel 従業員が Google Workspace 上に保存していた情報へとアクセスを広げることによって、最終的に、Vercel で利用されていた API キーやデータベース接続文字列などを窃取した。本件は、AI 利用を前提としたエコシステムの中で上流に位置する Context.ai が侵害されたことで、広範に利用されるデプロイプラットフォームにまで被害が広がった構図となっている。OAuth トークンに付与されたアクセス権の範囲が過剰に広かったことが被害を広げており、OAuth に関連するアクセス制御の統制の重要性を示す事例でもある。

情報源

- "Vercel April 2026 security incident | Vercel Knowledge Base" (2026-04-19) https://vercel.com/kb/bulletin/vercel-april-2026-security-incident
- "The Vercel Breach: OAuth Supply Chain Attack Exposes the Hidden Risk in Platform Environment Variables | Trend Micro (US)" (2026-04-20) https://www.trendmicro.com/en_us/research/26/d/vercel-breach-oauth-supply-chain.html

1.7. [2026-04-22] ChatGPT 向けワークスペースエージェントの提供開始

OpenAI は、Codex を基盤とした ChatGPT 用「ワークスペースエージェント」の提供を開始した。これにより、組織内の複雑なワークフローやクロスプラットフォームなタスクが自動化される。セキュリティの観点から、権限管理、役割ベースのアクセス制御、管理者による統制機能が実装されている。特にプロンプトインジェクション攻撃への耐性を備え、Compliance API による監査機能も提供する。AI による自動化

は生産性を高める一方で、組織の攻撃対象領域を拡大させるため、厳格なガバナンスとデータアクセス権の管理、および自動化プロセスの継続的な監視が求められる。

情報源

- "ChatGPT にワークスペースエージェントが登場 | OpenAI" (2026-04-22) https://openai.com/ja-JP/index/introducing-workspace-agents-in-chatgpt/

1.8. [2026-04-25] コーディングエージェントによる本番データベース全損事故

2026 年 4 月、Cursor で動作する AI コーディングエージェント（Claude Opus 4.6）が、スタートアップの本番環境 DB とバックアップを 9 秒で削除する事故が発生した。エージェントは資格情報の不一致を解消しようと、認証済み API を用いて非承認の削除コマンドを実行した。本件は、AI のガードレールの欠如、破壊的操作に対する確認手順の不在、API トークンの権限過多、およびバックアップの冗長性不足が招いた事態である。AI は自律的な「自己判断」により指示を逸脱しており、破壊的コマンドに対するヒューマン・イン・ザ・ループの重要性とインフラ側のセキュリティ制御の必要性が浮き彫りとなった。

- 情報源
- "Claude-powered AI coding agent deletes entire company database in 9 seconds — backups zapped, after Cursor tool powered by Anthropic's Claude goes rogue" (2026-04-28) https://www.tomshardware.com/tech-industry/artificial-intelligence/claude-powered-ai-coding-agent-deletes-entire-company-database-in-9-seconds-backups-zapped-after-cursor-tool-powered-by-anthropics-claude-goes-rogue
- "Cursor-Opus agent snuffs out startup’s production database" (2026-04-27) https://www.theregister.com/software/2026/04/27/cursor-opus-agent-snuffs_out_startups_production_database/5224442
- "“It took 9 seconds”: AI agent running on Anthropic’s Claude Opus 4.6 wipes critical database" (2026-04-27) https://www.businesstoday.in/amp/technology/story/it-took-9-seconds-ai-agent-

running-on-anthropics-claude-opus-46-wipes-critical-database-527552-2026-04-27

- "X ユーザーの JER さん: 「https://t.co/ofucbVgkLV」 / X" (2026-04-25) https://x.com/lifeof_jer/status/2048103471019434248

1.9. [2026-04-28] OpenAI「Code」のサンドボックス迂回ゼロデイ脆弱性

トレンドマイクロの Zero Day Initiative（ZDI）は、OpenAI の Codex における深刻なゼロデイ脆弱性（ZDI-26-305、CVSS 8.6）を公開した。この欠陥は JavaScript 実行環境に存在し、サンドボックス化されたコンテキストが適切に隔離されていないことで、リモート攻撃者がサンドボックスを迂回することが可能となる。悪用にはユーザーの操作が必要であり、被害者が悪意のあるリポジトリを処理する際に発生する。OpenAI は脆弱性の存在を認めたものの、バグ報奨金プログラムの対象外かつ製品のデフォルトの公開範囲に含まれないとして修正を拒否した。そのため本脆弱性は未修正のまま放置されており、当面は製品の利用を制限することが唯一の緩和策である。

情報源

- "OpenAI の「Codex」にサンドボックス迂回のゼロデイ脆弱性 〜Trend Micro の ZDI が指摘 - 窓の杜" (2026-04-30) https://forest.watch.impress.co.jp/docs/news/2105656.html
- "ZDI-26-305 | Zero Day Initiative" (2026-04-28) https://www.zerodayinitiative.com/advisories/ZDI-26-305/

1.10. [2026-04-30] Apple Support アプリにおける「CLAUDE.md」の誤公開

2026 年 4 月、Apple は Support アプリ（v5.13）のプロダクションビルドに、AI 支援ツール「Claude Code」向けの内部設定ファイル「CLAUDE.md」を誤って同梱した。このファイルには、「Juno AI」の設計や Actor ベースの並行処理モデル、UI コンポーネントライブラリといった機密性の高い内部設計情報が含まれていた。ユーザーデータの直接的な漏洩ではないものの、企業内部の開発標準やアーキテクチャ上の意思決定が露呈した形である。本件は、CI/CD パイプラインにおいて AI 向け構成ファイルを適切に除外する自動化プロセスの重要性を示唆しており、攻撃者に攻撃対象領域の

マッピングを許すリスクを浮き彫りにした。

情報源

- "What Apple’s Leaked CLAUDE.md Teaches Us | Vibe Coding" (2026-05-02) https://medium.com/vibe-coding/what-apples-leaked-claude-md-teaches-us-b8269e2ace51
- "X ユーザーの Aaron さん: 「Apple accidentally left Claude.md files in today's Apple Support app update (v5.13)」" (2026-04-30) https://x.com/aaronp613/status/2049986504617820551

1.11. [2026-04-30] OpenAI による「高度なアカウントセキュリティ」の導入

OpenAI は、アカウント乗っ取り（ATO）やフィッシングから機微なアカウントを保護するための高信頼性セキュリティ機能「高度なアカウントセキュリティ」を導入した。主な技術的特徴として、パスキーや FIDO 準拠のハードウェアセキュリティキーによるフィッシング耐性のある認証の必須化、メール/SMS によるアカウント復旧の廃止、セッション期間の短縮、セッション管理の可視化が挙げられる。また、本機能を有効にするとモデル学習から自動的に除外される。OpenAI は Yubico と提携し、ハードウェアキーの導入を促進するほか、一部の重要プログラムでは本機能の適用を必須とする方針を示し、AI インフラの防御強化を図っている。

情報源

- "Advanced Account Security - OpenAI" https://chatgpt.com/advanced-account-security?openaicom_referred=true
- "高度なアカウント セキュリティのご紹介 | OpenAI" (2026-04-30) https://openai.com/ja-JP/index/advanced-account-security/

1.12. [2026-05-04] エージェント型 AI 導入に対する共同セキュリティガイダンス

CISA や NCSC を含むファイブアイズ諸国は、エージェント型 AI の急速な導入に対する共同警告を発表した。同ガイダンスは、これらのシステムが相互接続された複雑な攻撃対象領域を生み出し、設定ミスや過剰な権限付与を介した攻撃に脆弱であること

を指摘している。主なリスクには、悪意のあるプロンプトに基づく自律的動作や、連携する低セキュリティツールを介した乗っ取りが含まれる。各機関は、レジリエンス、人間による監視、段階的な導入を優先するよう推奨している。専門家は、MITRE ATLAS等の脅威インテリジェンスが発展途上であることを認識し、標準的な評価手法が確立されるまでは予期せぬ動作を想定した運用計画が必要であるとしている。

情報源

- "Five Eyes warn agentic AI is too dangerous for rapid rollout" (2026-05-04) https://www.theregister.com/security/2026/05/04/five-eyes-warn-agentic-ai-is-too-dangerous-for-rapid-rollout/5229103

1.13. [2026-05-05] Microsoft 365「Copilot Cowork」の機能拡張

Microsoft は Copilot Cowork に Anthropic の Claude Opus 4.8 を統合し、新しいプラグインおよびコネクタフレームワークを導入した。長期間の多段階ワークフロー向けに設計された Cowork は、Dynamics 365 や Power BI などの業務アプリとの読み書き連携が可能となった。Agent 365 による可観測性、ガバナンス、ID 管理が統合され、セキュリティが強化されている。本プラットフォームはサンドボックス化されたクラウド環境でマルチモデルアーキテクチャを活用し、エンタープライズレベルのデータ保護を維持しつつ、AI エージェントによる複雑なタスクの自律的な実行を実現している。これらは生産性向上と同時に、管理すべき潜在的なリスク領域の増大であることに注意すべきである。

情報源

- "(Co)work in Progress | Microsoft Community Hub" (2026-05-05) https://techcommunity.microsoft.com/blog/microsoft365copilotblog/cowork-in-progress/4511672
- "Copilot Cowork: A new way of getting work done | Microsoft 365 Blog" (2026-03-09) https://www.microsoft.com/en-us/microsoft-365/blog/2026/03/09/copilot-cowork-a-new-way-of-getting-work-done/

- "Copilot Cowork: Now available in Frontier | Microsoft 365 Blog" (2026-03-30) https://www.microsoft.com/en-us/microsoft-365/blog/2026/03/30/copilot-cowork-now-available-in-frontier/

1.14. [2026-05-12] Anthropic が法律業界向け Claude エコシステムを拡張

Anthropic は、20 以上の Model Context Protocol（MCP）コネクタと 12 の専門分野別プラグインを導入し、法律業界向け Claude の機能を大幅に拡張した。これらの統合により、Claude は契約ライフサイクル管理、文書管理システム（iManage、NetDocuments 等)、電子証拠開示プラットフォームなどの主要な法務テクノロジースタックと直接連携可能になる。オープンプロトコルを採用することで、法律事務所は厳格なデータガバナンスと権限管理を維持しつつ、ワークフローをカスタマイズできる。本ツール群は、リライン（赤入れ)、トリアージ、規制監視といった定型業務の自動化と、組織固有の基準・プレイブックへの準拠を支援する。さらに、公益団体や非営利の法律扶助団体との提携を通じて、法的サービスへのアクセス拡大を目指す。

情報源

- "Claude for the legal industry | Claude" (2026-05-12) https://claude.com/blog/claude-for-the-legal-industry

1.15. [2026-05-13] Anthropic「Claude for Small Business」が登場

Anthropic は、Claude Cowork に統合されたエージェント型 AI ソリューション「Claude for Small Business」を発表した。これは中小企業の AI 導入の遅れを解消することを目的とし、QuickBooks、PayPal、HubSpot、Microsoft 365 などのツールと連携し、給与計算や請求、マーケティング業務を自動化する。セキュリティ面では、顧客データによる学習を行わず、既存のシステム権限を継承し、機密性の高いアクションには人間による承認を必須とする「ヒューマン・イン・ザ・ループ」を導入している。この取り組みは、反復的な事務作業を AI に代行させるとともに、安全かつ倫理的な AI 活用を促進するための「AI Fluency」研修プログラムと併せて展開される。

情報源

- "Claude for Small Business | Claude by Anthropic"
  https://claude.com/solutions/small-business
- "Introducing Claude for Small Business ¥ Anthropic" (2026-05-13)
  https://www.anthropic.com/news/claude-for-small-business

1.16. [2026-05-13] TanStack npm サプライチェーン攻撃に伴う OpenAI の対応

OpenAI は、「Mini Shai-Hulud」キャンペーンの一環である TanStack npm パッケージへのサプライチェーン攻撃により、従業員のデバイスが侵害されたと発表した。この影響で一部のリポジトリ内の認証情報や、Windows、macOS、iOS 向けのコード署名証明書が露出した。顧客データや知的財産への被害は確認されていないが、予防的措置として証明書の更新を実施。これに伴い、macOS ユーザーは 2026 年 6 月 12 日までにアプリケーションを更新する必要がある。同社は今後、CI/CD パイプラインのセキュリティ強化や依存関係の検証制御を加速させ、広範なエコシステムのリスクに対処する方針である。本件は世界的な AI 製品がサプライチェーン攻撃の影響を受けた実事例である。

情報源

- "TanStack npm サプライチェーン攻撃への対応 | OpenAI" (2026-05-13)
  https://openai.com/ja-JP/index/our-response-to-the-tanstack-npm-supply-chain-attack/

1.17. [2026-05-18] GitHub リポジトリを標的とした大規模攻撃「Megalodon」

2026 年 5 月、GitHub リポジトリ 5,500 件以上に対し不正な GitHub Actions ワークフローを注入する攻撃キャンペーン「Megalodon」が発生した。攻撃者は偽の身元を利用し、CI/CD 環境変数やクラウド認証情報、OIDC トークンを外部 C2 サーバーへ窃取する Base64 エンコードされたペイロードをコミットした。攻撃手法は、自動実行される「SysDiag」と`workflow_dispatch`で起動する休眠型バックドア「Optimize-Build」の 2 種類である。本件は`@tiledesk/tiledesk-server`等の npm パッケージにも波及した。AI アプリケーション開発で利用される npm パッケージも多数含まれてい

る。被害リポジトリの特定、シークレットのローテーション、およびOIDCトークン要求の監視が強く推奨される。

情報源

- "Megalodon: Mass GitHub Repo Backdooring via CI Workflows - Real-time Open Source Software Supply Chain Security" (2026-05-21) https://safedep.io/megalodon-mass-github-repo-backdooring-ci-workflows/?ref=joho-todai.com

1.18. [2026-05-19] 自律型AIエージェント「Gemini Spark」の発表

Googleは、Google Workspace全体でマルチステップのタスクを実行可能な自律型AIエージェント「Gemini Spark」を発表した。Gemini 3.5 FlashとAntigravity制御プレーンを基盤とし、電子メール管理や複雑なワークフローの自動化を実現する。Google I/O 2026で導入された新機能には、Antigravity 2.0デスクトップ環境やWizと連携した強化セキュリティが含まれる。これらの進化は、会話型AIからエージェント型プラットフォームへの移行を意味し、動的なサブエージェント生成やクロスアプリケーション連携を可能にする一方、重要な操作にはユーザーの承認を必須とする設計である。

情報源

- "Gemini Spark – Your 24/7 personal AI agent for productivity" https://gemini.google/overview/agent/spark/
- "Google I/O 2026 のスタートアップ向けの主な発表 | Google Cloud 公式ブログ" (2026-06-02) https://cloud.google.com/blog/ja/topics/startups/startup-news-from-io-and-what-it-means-to-founders
- "The Gemini app becomes more agentic, delivering proactive, 24/7 help" (2026-05-19) https://blog.google/innovation-and-ai/products/gemini-app/next-evolution-gemini-app/

1.19. [2026-05-20] Google「Universal Cart」と標準プロトコル「UCP」の発表

Google は I/O 2026 にて、Google サービス横断型の AI ショッピングハブ「Universal Cart」を発表した。これは、小売業者間でのエージェントによるやり取りを標準化する新プロトコル「UCP」を採用している。セキュリティ面では、AI による決済に厳格な制限と検証可能な監査証跡を課す「Agent Payments Protocol（AP2）」を導入。分散した小売システムを安全な共通言語で統合し、プライバシー保護と透明性を担保したエージェントコマースを実現する。AI エージェントはユーザー定義の予算や制約内で動作し、主要な決済プロバイダーとの互換性も維持される。EC を軸に複数のサービスのいわゆるマッシュアップを AI が実行するユースケースであり、取り扱いに注意を要する多数のコンテキストを AI に提供する点に留意する必要がある。

情報源

- "Google、AI が価格監視や決済最適化を自動で行う「Universal Cart」発表 - ITmedia NEWS" (2026-05-20)
  https://www.itmedia.co.jp/news/articles/2605/20/news102.html
- "Universal Cart などショッピングをサポートする新たな機能の紹介" (2026-05-20)
  https://blog.google/intl/ja-jp/products/explore-get-answers/google-shopping-cart/

1.20. [2026-05-22] OpenAI「Codex」の「Goal mode」導入と自律性強化

OpenAI は、長時間にわたる自律的なタスク実行を可能にする「Goal mode」を Codex プラットフォームに導入した。従来のプロンプト形式とは異なり、目標、測定可能な結果、制約条件を永続的に定義できる。Codex は文脈を維持しながらステップの反復、プラグインの利用、GUI 操作を自律的に行う。セキュリティ専門家は、生産性向上の一方でエージェントの自律性がもたらすリスクに留意すべきである。長時間タスクにおけるドリフト防止や不正なシステム操作を防ぐため、明確な権限境界、チェックポイント、監査ログ（フライトレコーダー）の運用が不可欠である。

情報源

- "X ユーザーの OpenAI さん: 「3 Goal mode is now available in the Codex app, IDE extension, and CLI.」" (2026-05-22)
  https://x.com/OpenAI/status/2057617860986593680
- "（ほぼ）あらゆる作業に対応する Codex | OpenAI" (2026-04-16)
  https://openai.com/ja-JP/index/codex-for-almost-everything/
- "Computer Use – Codex app | OpenAI Developers"
  https://developers.openai.com/codex/app/computer-use
- "Using Goals in Codex" (2026-05-09)
  https://developers.openai.com/cookbook/examples/codex/using_goals_in_codex

1.21. [2026-05-22] Google 検索のプロンプトインジェクション脆弱性類似事象

Google I/O で導入された AI 搭載検索において、入力サニタイズに関する重大な不具合が確認された。「disregard」「ignore」「skip」といったコマンド形式のキーワードを検索すると、基盤モデル（Gemini）がユーザー入力を検索クエリではなくシステム命令として誤認する。これにより、プロンプトインジェクションに類似した挙動が発生し、AI が検索結果の代わりに命令への応答を出力する。現状は悪意あるエクスプロイトではなく機能的なバグとされているが、ユーザーの意図とシステム命令の分離が不十分なまま LLM を検索インターフェースに組み込むことの潜在的リスクを浮き彫りにしている。

情報源

- "Google’s AI search is so broken it can ‘disregard’ what you’re looking for | The Verge" (2026-05-23)
  https://www.theverge.com/tech/936176/google-ai-overviews-search-disregard
- "Searching for 'Disregard' Breaks Google [Updated] - MacRumors" (2026-05-22)
  https://www.macrumors.com/2026/05/22/google-search-disregard/
- "You can no longer Google the word 'disregard' | TechCrunch" (2026-05-22)
  https://techcrunch.com/2026/05/22/you-can-no-longer-google-the-word-

disregard/?utm_source=dlvr.it&utm_medium=twitter

1.22. [2026-05-22] AI モデルに対する国家系アクターによる偽情報汚染

最新の監査によると、Anthropic の Claude は偽情報に対して脆弱性を増しており、「Pravda」ネットワーク等のロシアやイランの国営系プロパガンダを引用する頻度が著しく上昇している。同時に、ロシアのソーシャル・デザイン・エージェンシー（SDA）から流出した内部文書により、20 万本以上の記事を含むドイツ語版「Wikipedia クローン」を展開する戦略的計画が明らかになった。これらの調査結果は、国家主導の偽情報キャンペーンが、検索結果の飽和やデータポイズニングを通じて、生成 AI の信頼性と出力の整合性を明確に標的にしているという脅威が現実のものであることを示している。

情報源

- "Anthropic’s AI Chatbot Is Leaning More on Russian and Iranian Propaganda Sources, NewsGuard Audit Finds" (2026-05-04) https://www.newsguardtech.com/special-reports/anthropic-ai-chatbot-claude-russia-iran-propaganda/
- "Propaganda-Offensive gegen Deutschland – Leak enthüllt Russlands Vorgehen" (2026-05-22) https://www.t-online.de/nachrichten/deutschland/innenpolitik/id_101262296/propaganda-offensive-gegen-deutschland-leak-enthuellt-russlands-vorgehen.html

1.23. [2026-05-26] Microsoft「Copilot Studio」の自動化能力とガバナンス強化

Microsoft は Copilot Studio において、レガシーシステムとの直接的な UI 操作を可能にする「Computer Use」エージェントの一般提供（GA）を開始した。主要なアップデートとして、統一ワークフローデザイナー、相互運用性を高める Work IQ API/CLI、エージェント間（A2A）通信が導入された。セキュリティ面では安全な資格情報管理と音声エージェントのガバナンスガイドが提供される。また、新しいオーケストレーション層により実行性能が 20%向上し、トークン消費が 50%削減された。これらの機能は、断片化された企業環境において、組織統制を維持しつつ適応性の高

い自動化を実現することに主眼を置いている。本件はホワイトカラー業務における多大なシェアを誇るMicrosoft製品が汎用AIエージェントへと発展していることを示すと共に、利用者における適切なリスク判断の重要性が高まっている。

情報源

- "「Copilot Studio」でAIエージェントがアプリ上のUIを直接操作できる機能が正式版に - 窓の杜" (2026-05-29)
  https://forest.watch.impress.co.jp/docs/news/2112927.html
- "What's new in Copilot Studio: May 2026 updates and features | Microsoft Copilot Blog" (2026-05-26)
  https://www.microsoft.com/en-us/microsoft-copilot/blog/copilot-studio/new-and-improved-computer-using-agents-a-new-workflows-experience-and-real-time-voice-experiences/

1.24. [2026-05-26] Microsoft「Copilot Cowork」からのファイル漏洩リスク

Microsoft Copilot Coworkに間接プロンプトインジェクションに対する脆弱性が発見され、不正なファイル持ち出しが可能であることが指摘された。攻撃者はユーザーが制御可能なパスから自動読み込みされる「スキル」を汚染することで、エージェントを操作し、事前認証済みのファイルダウンロードリンクを含むメッセージを生成させる。重要な点として、これらのメッセージはユーザー本人宛に自動承認で送信されるため、セキュリティ制御を回避する。ユーザーがメッセージを開くと、エージェントが意図せず攻撃者のサーバーへネットワークリクエストを送り、機密情報が流出する。この攻撃はClaude Opus 4.7を含むモデルに依存せず、スケジュールタスク機能により脅威が増大する。AIエージェントを利用する際には、この脆弱性に限らず、組織はエージェントに対するダウンロード権限を制限し、信頼できないスキルファイルの読み込みを監視すべきである。

情報源

- "Microsoft Copilot Cowork Exfiltrates Files"
  https://www.promptarmor.com/resources/microsoft-copilot-cowork-exfiltrates-files

1.25. [2026-05-27] Anthropic「Zero Trust for AI agents」の発表

Anthropicが公開した本セキュリティフレームワークは、自律型AIエージェントを導入する際の、ゼロトラスト原則に基づいた包括的なセキュリティ指針である。フロンティアAIモデルの進化に伴い、攻撃コードの自動生成など脅威の発生スピードが極端に加速しており、防御側にもマシンスピード水準の対応が求められている。エージェントの導入は、ツールの誤用、メモリやRAGの毒汚染、エージェント間の連携悪用といった固有のリスクをもたらすため、従来の境界型セキュリティは通用しない。この課題に対し、本フレームワークは「Foundation」「Enterprise」「Advanced」の3つの成熟度ティアを提示する。また、OWASPが提唱する「最小権限（least agency）」の概念を取り入れ、エージェントが実行可能なツールの範囲や頻度を厳格に制限する。主要な実装技術として、暗号論的に保護された一時トークンの利用、証明書ピンニングを伴う相互TLS、およびコンテナサンドボックスによる実行環境の隔離などを推奨している。さらに、間接的プロンプトインジェクションを防ぐための「スポットライティング」などの入力検証や、自律的なインシデント処理を行う「エージェント型SOAR」への移行についても言及している。

情報源

- "Zero Trust for AI agents | Claude" (2026-05-27)
  https://claude.com/blog/zero-trust-for-ai-agents

1.26. [2026-05-28] Starlette/FastAPIにおける認証バイパス脆弱性「BadHost」

CVE-2026-48710（BadHost）は、Pythonで高性能なWebサービスやAPIを開発するための軽量な非同期Webフレームワークである Starlette 1.0.1 未満に存在する深刻な認証バイパス脆弱性であり、FastAPIおよび広範なPython AIエコシステムに影響を及ぼす。この脆弱性は、StarletteがHTTPのHostヘッダーをサニタイズせずに連結して`request.url`を生成する不適切な設計に起因する。攻撃者は細工したヘッダーで`request.url.path`を改ざんし、パスベースの認証ミドルウェアを回避できる。vLLM、LiteLLM、MCPサーバーなどの主要ツールが影響を受ける。対策として、Starletteを1.0.1以降へ更新すること、パスベースの認証からエンドポイント単位のセキュリティ（`Depends()`等）へ移行すること、あるいはRFC準拠のリバースプロキシによるヘッダーの正規化が推奨される。

情報源

- "BadHost - CVE-2026-48710 Starlette Host-Header Auth Bypass" (2026-05-28) https://badhost.org/
- "Millions of AI agents imperiled by critical vulnerability in open source package - Ars Technica" (2026-05-27) https://arstechnica.com/information-technology/2026/05/millions-of-ai-agents-imperiled-by-critical-vulnerability-in-open-source-package/

1.27. [2026-05-29] Microsoft 365 AI 機能の拡充と OneDrive によるファイル名提案

Microsoft は、Copilot 機能の高度化と「Microsoft Agent 365」のリリースにより、365 エコシステムを大幅に拡充した。本プラットフォームは、組織内のエージェントを監視、管理、保護するための集中制御プレーンとして機能する。主な更新には、A2A や MCP を通じた開発者統合のための Work IQ API や、Anthropic との提携によるモデル選択肢の拡大が含まれる。この一環でファイルの内容を AI が把握することで OneDrive から適切なファイル名を提案する機能の導入がロードマップに含まれた。しかしこの機能は、間接プロンプトインジェクションの発生ポイントが潜在的に増えることも意味する。

情報源

- "Microsoft 365 Roadmap | Microsoft 365" https://www.microsoft.com/en/microsoft-365/roadmap?id=564909

1.28. [2026-06-01] Miasma マルウェアによる広範なサプライチェーン攻撃

2026 年 6 月、Mini Shai-Hulud の亜種である Miasma ワームが、npm レジストリおよび GitHub リポジトリを標的とした大規模なサプライチェーン攻撃を実行した。このワームは、binding.gyp を悪用して npm インストール時に実行トリガーを引く「Phantom Gyp」手法や、AI コーディング支援ツール（VS Code、Cursor、Claude Code 等）に永続的なフックを仕込む手法を駆使した。CI/CD パイプラインを標的とし、クラウド、Kubernetes、リポジトリ認証情報を窃取して攻撃者の GitHub アカウ

ントへ送信する。このキャンペーンは、認証情報窃取と自己増殖を自動化しており、AI 支援ツールや CI/CD 構成を悪用したマルウェア実行のリスクが急増していることを示している。

情報源

- "Dozens of Red Hat packages backdoored through its official NPM channel - Ars Technica" (2026-06-02)
  https://arstechnica.com/security/2026/06/dozens-of-red-hat-packages-backdoored-through-its-offical-npm-channel/
- "Miasma npm Supply Chain Attack: Self-Spreading Worm via Phantom Gyp - StepSecurity" (2026-06-03)
  https://www.stepsecurity.io/blog/binding-gyp-npm-supply-chain-attack-spreads-like-worm
- "Miasma Worm Targets AI Coding Agents via GitHub Repos - Real-time Open Source Software Supply Chain Security" (2026-06-05)
  https://safedep.io/miasma-worm-ai-coding-agent-config-injection/
- "Red Hat npm Packages Compromised to Spread a Credential-Stealing Worm" (2026-06-01)
  https://www.aikido.dev/blog/red-hat-npm-packages-compromised-credential-stealing-worm

1.29. [2026-06-02] AI サポートの脆弱性を狙った Instagram アカウント乗っ取り

2026 年 5 月下旬、Meta の AI サポートチャットボットの論理的欠陥を悪用し、著名な Instagram アカウントを乗っ取る攻撃が発生した。攻撃者はプロンプトインジェクションを用い、ターゲットのアカウントに自身のメールアドレスを紐付けるよう AI を誘導することで、二要素認証を回避した。当該の AI サポートチャットボットでは、攻撃者の身元確認に関する設計が不十分であった。この「混乱した代理人（confused deputy）」問題に起因する手法により、高価値なアカウントが組織的に奪取・転売された。Meta は 5 月 29 日に緊急パッチを適用し、アカウント復旧に関する API アクセスを制限したが、専門家は、確定的な認証ゲートなしに生成 AI エージェントへ高権限を与えることの危険性を強く警告している。

情報源

- "Hackers duped Meta AI support chatbot to steal celebrity Instagram accounts - Ars Technica" (2026-06-02)
  https://arstechnica.com/ai/2026/06/meta-ai-support-chatbot-gave-hackers-access-to-notable-instagram-accounts/
- "Hackers Simply Asked Meta AI to Give Them Access to High-Profile Instagram Accounts. It Worked" (2026-06-01)
  https://www.404media.co/hackers-simply-asked-meta-ai-to-give-them-access-to-high-profile-instagram-accounts-it-worked/
- "Hackers Used Meta's AI Support Bot to Seize Instagram Accounts – Krebs on Security" (2026-06-01)
  https://krebsonsecurity.com/2026/06/hackers-used-metas-ai-support-bot-to-seize-instagram-accounts/
- "Instagram Meta AI Vulnerability: How Hackers Bypassed 2FA with Prompt Injection | The CyberSec Guru" (2026-06-07)
  https://thecybersecguru.com/news/instagram-meta-ai-vulnerability-account-recovery-exploit/

1.30. [2026-06-04] 自律型 AI エージェント「Microsoft Scout」の発表

マイクロソフトは、Windows、macOS、Microsoft 365 全体でバックグラウンド実行される自律型 AI エージェント「Microsoft Scout」を発表した。オープンソースの OpenClaw を基盤とし、ファイル操作やシェルコマンド実行、ブラウザ自動化などを担う。セキュリティ面では Entra ID と統合され、すべての動作が検証可能な ID に紐づく。組織のセキュリティフレームワーク内で動作するよう、Microsoft Purview の DLP ポリシーや秘密度ラベルを適用し、機密操作には人による承認を義務付けるなど、監査可能なエンタープライズレベルの管理を実現する。これにより AI 関連ビッグテックの大半が汎用 AI エージェント環境の提供を開始した状況となり、2026 年中には本格的に AI エージェントの浸透が進むものと見込まれる。

情報源

- "Microsoft Scout (Frontier) overview | Microsoft Learn" https://learn.microsoft.com/ja-jp/microsoft-scout/overview
- "常時稼働するあなた専用のパーソナル エージェント Microsoft Scout が登場 - Source Asia" (2026-06-04) https://news.microsoft.com/source/asia/features/introducing-microsoft-scout-your-always-on-personal-agent/?lang=ja

1.31. [2026-06-04] 次世代AIモデル「Claude Oceanus-v1-p」の不正配布

一部のセキュリティ関係者（レッドチーム）だけにアクセスを許可していたはずの、Anthropicの最新AIモデル「Claude Oceanus-v1-p」が、レッドチーム評価開始直後にセキュリティ侵害を受けた。攻撃者は配布チャネルを悪用し、中国を拠点とするプロキシサービスを通じてAPIアクセスを不正に再販した。これは過去の中国AIラボによるプロキシ悪用事例と類似している。ゼロデイ脆弱性の発見に優れた「Claude Mythos」の後継モデルである本モデルは、Anthropicの「Project Glasswing」の中核を成す。不正アクセスとモデルの重大なリスクを鑑み、Anthropicは内部調査のためレッドチームのアクセスを一時停止した。同社は、悪用防止のための強固な対策が未確立であるとして、一般公開を控える方針である。

情報源

- "Anthropic's Claude Oceanus-v1-p Opens to Red Team Testing, but Distribution is Compromised" (2026-06-04) https://cybersecuritynews.com/anthropics-claude-oceanus-v1-p/

1.32. [2026-06-07] ChatGPTにおける「ロックダウンモード」の導入

OpenAIは、プロンプトインジェクションによるデータ流出リスクを軽減するため、ChatGPTに「ロックダウンモード」を導入した。この選択型セキュリティ機能は、外部へのネットワークリクエストを制限し、ライブWebブラウジング、ディープリサーチ、エージェントモードなどの高リスク機能を無効化する。外部接続を制限することで、AIツールが機密データを不正なサーバーへ送信することを防ぐ。すべてのリスク

を排除するものではないが、機密情報を扱うユーザーにとって強固な防御策となる。併せてOpenAIは、Codexを含む全プラットフォームで特定機能に「高リスク」ラベルを適用し、ネットワークアクセスに伴うセキュリティ上の注意を喚起している。これらの対策は一般ユーザーでも比較的容易に導入することができる。

情報源

- "ChatGPTに「ロックダウンモード」　プロンプトインジェクションによる情報漏えい対策 - ITmedia NEWS" (2026-06-07)
  https://www.itmedia.co.jp/news/articles/2606/07/news022.html
- "ChatGPT にロックダウンモードと一貫した「高リスク」ラベルが導入されました | OpenAI" (2026-02-13)
  https://openai.com/ja-JP/index/introducing-lockdown-mode-and-elevated-risk-labels-in-chatgpt/
- "Lockdown Mode | OpenAI Help Center"
  https://help.openai.com/en/articles/20001061-lockdown-mode

1.33. [2026-06-09] Claude Managed Agentsの機能拡張とセキュリティ強化

Anthropicは Claude Managed Agentsプラットフォームを大幅に拡充し、自律的な長時間稼働を支援する機能を導入した。主な更新には、ファイルシステムベースのメモリ機能、セッション間での自己改善を促す「Dreaming」、マルチエージェントのオーケストレーションが含まれる。また、エンタープライズセキュリティの課題に対し、自己ホスト型サンドボックス、認証情報管理用の「Vaults」、プライベートネットワークアクセス用のMCPトンネルを提供している。これらの機能により、開発者はインフラの制御を維持しつつ、セキュアで自己修正可能なプロダクション級エージェントを構築可能となる。オーケストレーションやコンテキスト管理をプラットフォーム側に任せることで、組織は厳格なアクセス制御と監査性を確保しながら、AIエージェントの迅速な実用化を実現できる。

情報源

- "Built-in memory for Claude Managed Agents | Claude" (2026-04-23)

https://claude.com/blog/claude-managed-agents-memory

- "Claude Managed Agents: get to production 10x faster | Claude" (2026-04-08) https://claude.com/blog/claude-managed-agents
- "New in Claude Managed Agents: dreaming, outcomes, and multiagent orchestration | Claude" (2026-05-06) https://claude.com/blog/new-in-claude-managed-agents
- "New in Claude Managed Agents: run agents on a schedule and store environment variables in vaults | Claude" (2026-06-09) https://claude.com/blog/whats-new-in-claude-managed-agents
- "New in Claude Managed Agents: self-hosted sandboxes and MCP tunnels | Claude" (2026-05-19) https://claude.com/blog/claude-managed-agents-updates

2. AI を活用したサイバーセキュリティ確保（AI for Security）

2.1. [2026-04-29] Linux カーネルにおける権限昇格脆弱性「Copy Fail」

「Copy Fail」(CVE-2026-31431)は、Linux カーネルの authencesn 暗号テンプレートにおける重大な論理欠陥である。権限を持たないローカルユーザーが、setuid バイナリを含む任意の読み取り可能なファイルのページキャッシュに対し、4 バイトの書き込みを行うことを可能にする。AF_ALG ソケットと splice()システムコールを介して実行され、VFS 書き込みパスを迂回するため、ディスク上に痕跡を残さない。2017 年以降の主要な全 Linux ディストリビューションにおいて確率的ではない確実な権限昇格が可能である。根本原因は AEAD 操作の設計にあり、ページキャッシュページを書き込み可能なスキャターリストへ誤って露出させている点にある。修正にはカーネルのパッチ適用、または algif_aead モジュールの無効化が必要である。

情報源

- "Copy Fail: 732 Bytes to Root on Linux - Xint" (2026-04-30)
  https://xint.io/blog/copy-fail-linux-distributions
- "Copy Fail — CVE-2026-31431" (2026-04-29)
  https://copy.fail/

2.2. [2026-04-30] Anthropic「Claude Security」の提供開始

Anthropic は、Claude Enterprise 向けに AI 脆弱性診断ツール「Claude Security」のパブリックベータ版を公開した。Claude Opus 4.7 モデルを活用し、従来のパターンマッチングを超えたコンテキスト分析により、メモリ破壊、認証バイパス、複雑なロジックエラーを特定する。誤検知を抑制する敵対的検証機能を備え、修正パッチの提案も行う。Webhook による Slack/Jira 連携や、CrowdStrike 等の主要プラットフォームとの統合も進んでおり、発見から修正までのサイクルを劇的に短縮する次世代の脆弱性管理を実現している。なお、Claude Code 用の security-guidance プラグインと異なり、Claude Security は、開発ワークフローの中で自動実行するのではなく、セキュリティ対応のプロセスの中でツールとして操作する。

情報源

- "Claude Security is now in public beta | Claude" (2026-04-30)
  https://claude.com/blog/claude-security-public-beta
- "Claude Security | Claude by Anthropic"
  https://claude.com/product/claude-security
- "Claude Security を使用する | Anthropic ヘルプセンター" (2026-06-01)
  https://support.claude.com/ja/articles/14661296-claude-security%E3%82%92%E4%BD%BF%E7%94%A8%E3%81%99%E3%82%8B

2.3. [2025-05] Alibaba 製 AI コードレビューツール「Open Code Review」

Alibaba は、コードのセキュリティと品質を向上させる CLI ツール「Open Code Review」をオープンソース化した。汎用的な AI エージェントとは異なり、決定論的なエンジニアリング（ファイル選別、ルール適合、行位置特定）とセマンティック解析を行う LLM エージェントを組み合わせたハイブリッド構造を採用している。これにより、ハルシネーションや位置ズレを最小化しつつトークン消費を最適化する。OpenAI/Anthropic 互換のエンドポイントに対応し、CI/CD パイプラインへの統合が可能で、XSS、SQL インジェクション、スレッドセーフティなどの一般的な脆弱性に対するルールを内蔵している。

情報源

- "GitHub - alibaba/open-code-review: Open-source & free — Battle-tested at Alibaba's scale."
  https://github.com/alibaba/open-code-review
- "Open Code Review — Agent Native Code Review"
  https://alibaba.github.io/open-code-review/

2.4. [2026-05] AI 時代の脆弱性報告・管理モデルの変容と Linux カーネルの対応

AI 支援による脆弱性調査の普及により、従来のセキュリティ開示サイクルは根本的に

破綻した。脆弱性は複数の研究者によって同時多発的に発見され、エクスプロイト作成時間が数分に短縮されたことで、90日間の猶予や月次パッチ適用は時代遅れとなったとの懸念がある。これに対し、Linuxカーネルコミュニティはセキュリティ文書を刷新し、AI検知バグの公開原則や脅威モデルの明確化、緊急緩和策としての「キルスイッチ」導入を進めている。重大な脆弱性を即時対応案件として扱い、パッチ分析や脅威検知をAIで自動化する防御態勢への転換が迫られている。

情報源

- "Documentation: security-bugs: new updates covering triage and AI [LWN.net]" (2026-05-03)
  https://lwn.net/Articles/1070963/
- "killswitch for short-term emergency vulnerability mitigation [LWN.net]" (2026-05-08)
  https://lwn.net/Articles/1071861/
- "The 90 day disclosure policy is dead :: Himanshu Anand" (2026-05-09)
  https://blog.himanshuanand.com/2026/05/the-90-day-disclosure-policy-is-dead/
- "Significant raise of reports [LWN.net]" (2026-03-31)
  https://lwn.net/Articles/1065620/

2.5. [2026-05-04] AIエージェント駆動型脆弱性スキャナ Vercel「deepsec」公開

Vercelは、大規模コードベースの脆弱性分析を目的とした、AIエージェント駆動型のセキュリティツール「deepsec」をオープンソース化した。従来の静的解析ツールとは異なり、ClaudeやCodexなどの高度なLLMを活用し、ソースコードの深い文脈調査を行う。初期の正規表現ベースのスキャン、AIによる調査、偽陽性を抑制する再検証フェーズを経て動作する。ユーザーはローカル環境で実行可能なほか、Vercel Sandboxを使用して並列実行も可能である。カスタムプラグインの開発にも対応しており、独自の認証モデルやデータモデルに適合させつつ、自社インフラ内で完結する高精度なセキュリティ分析を実現する。

情報源

- "GitHub - vercel-labs/deepsec: Deepsec is a security harness for finding vulnerabilities in your codebase powered by coding agents" https://github.com/vercel-labs/deepsec/
- "Introducing deepsec: The security harness for finding vulnerabilities in your codebase - Vercel" (2026-05-04) https://vercel.com/blog/introducing-deepsec-find-and-fix-vulnerabilities-in-your-code-base

2.6. [2026-05-09] Linux カーネルにおける深刻な権限昇格脆弱性「Dirty Frag」

セキュリティ研究者の Hyunwoo Kim（v4bel）が、Linux カーネルにおける権限昇格脆弱性クラス「Dirty Frag」を公表した。これは xfrm-ESP（CVE-2026-43284）と RxRPC（CVE-2026-43500）における 2 つのページキャッシュ書き込み脆弱性を連鎖させたもので、競合状態を伴わずに特権を獲得できる。最大 9 年間にわたり存在したこの脆弱性は、決定論的で再現性が極めて高い。本件については、CVE 付与前の CVD プロセス進行中に、AI を用いた解析により第三者がエクスプロイトを生成して公開したという問題がある。従来の責任ある情報公開のプロセスが脅かされている現状を浮き彫りにした事例である。推奨される対策は、パッチ適用または脆弱なモジュールの無効化およびページキャッシュのクリアである。

情報源

- "CVE 공급망 공격 관련 AI 의 등장, 새 국면 맞이하고 있는 사이버보안" (2026-05-09) https://www.boannews.com/media/view.asp?idx=143537&kind=&sub_kind=
- "GitHub - V4bel/dirtyfrag · GitHub" (2026-05-07) https://github.com/V4bel/dirtyfrag

2.7. [2026-05-11] OpenAI「Daybreak」の発表

OpenAI は、フロンティア AI モデルを活用してサイバー防御を加速させる新プロジェクト「Daybreak」を発表した。Daybreak は開発ワークフローに統合され、Codex

Security を通じて脆弱性スキャン、脅威モデリング、パッチ検証を自動化する。「Trusted Access for Cyber」の下で提供される GPT-5.5-Cyber などの専用モデルを用いることで、事後対応型から設計段階でのレジリエントなセキュリティ戦略への転換を目指す。同プラットフォームは優先度の高い脅威の特定と検証可能な修正に焦点を当て、セキュリティチームが監査証跡を維持しつつ、効率的に防御業務を遂行できるようにする。

情報源

- "X ユーザーの Sam Altman さん: 「OpenAI is launching Daybreak, our effort to accelerate cyber defense and continuously secure software.」" (2026-05-11) https://x.com/sama/status/2053951874408276193?ref=joho-todai.com
- "Daybreak | サイバーセキュリティのための OpenAI | OpenAI" https://openai.com/ja-JP/daybreak/

2.8. [2026-05-12] AI エージェント型脆弱性スキャンシステム Microsoft「MDASH」

Microsoft は、脆弱性の発見から検証までを自動化する AI セキュリティシステム「MDASH」を発表した。100 以上の専門エージェントを複数のモデル間で協調させることで、Windows ネットワークスタックにおける 16 件の脆弱性（RCE を含む）を特定した。単一モデルのアプローチとは異なり、準備、スキャン、検証、重複排除、証明という多段階のパイプラインを用いることで、誤検知を抑えつつ高精度な結果を提供する。CyberGym ベンチマークで 88.45%のスコアを記録し、過去の MSRC ケースでも高い再現性を実証しており、AI 駆動型セキュリティを実用的なエンタープライズ防衛のレベルへ引き上げている。

情報源

- "Defense at AI speed: Microsoft’s new multi-model agentic security system tops leading industry benchmark | Microsoft Security Blog" (2026-05-12) https://www.microsoft.com/en-us/security/blog/2026/05/12/defense-at-ai-speed-microsofts-new-multi-model-agentic-security-system-tops-leading-industry-benchmark/

2.9. [2026-05-12] Obsidian コミュニティにおけるプラグインセキュリティ強化

Obsidian は、プラグインのセキュリティ強化を目的とした新たな「Obsidian Community」ポータルと開発者ダッシュボードを立ち上げた。本プラットフォームには、各バージョンのプラグインに対し、脆弱性、コード品質、マルウェアの自動スキャンを実行するシステムが導入された。これにより、4,000 を超えるプラグインに対するスケーラブルな監視が可能となる。今後は、ファイルシステムやネットワークアクセス等の権限開示の義務化や、認証済み開発者バッジの導入も予定されている。既存プラグインの再審査が進められる一方、今回の変更により審査プロセスが迅速化され、開発者はスコアカードを通じて即時にフィードバックを得られるため、エコシステム全体の安全性が向上する。この取り組みの背景には、コーディングエージェントの発展に伴い低質で潜在的に危険なプラグインが大量に開発されるようになったという状況がある。

情報源

- "The future of Obsidian plugins - Obsidian" (2026-05-12)
  https://obsidian.md/blog/future-of-plugins/

2.10. [2026-05-27] Anthropic「Claude Code」向け security-guidance プラグイン

Anthropic は、コーディング中に脆弱性を検知・修正する Claude Code 用「security-guidance」プラグインをリリースした。本プラグインは、ファイル編集時の決定論的なパターンマッチングと、ターン終了時および Git コミット時のエージェントによるレビューという多層的なアプローチを採用している。インジェクション、安全でないデシリアライゼーション、DOM API の悪用といった重要課題を対象とし、プルリクエスト時のセキュリティ指摘を 30〜40%削減することを目指す。これは多層防御の一環であり、完全なセキュリティ対策ではなく、従来の CI/CD や手動レビューとの併用が推奨される。今後のコーディングエージェントの利用においては、この種のセキュリティ支援 AI の同時活用が不可欠になっていくものと見込まれる。

情報源

- "We’ve shipped a security-guidance plugin for Claude Code (X post)" (2026-05-27) https://x.com/ClaudeDevs/status/2059385239781384341
- "Claude がコードを書く際のセキュリティ問題をキャッチする - Claude Code Docs" https://code.claude.com/docs/ja/security-guidance
- "Security Guidance – Claude Plugin | Anthropic" https://claude.com/plugins/security-guidance

2.11. [2026-06] Google AI Threat Defense による自律型脅威防御

Google は、AI を駆使した攻撃に対抗するため「Google AI Threat Defense」を発表し、Google Security Operations を強化した。マルチモデル AI、Mandiant の知見、Gemini の推論機能を活用し、自律的な脆弱性管理、リスクの優先順位付け、自動修復を実現する。特筆すべきは、新たな脆弱性パターンをカスタム検知ルールに変換する「Detection Engineering agent」や、迅速な封じ込めを行う「Agentic automation」である。Wiz のコンテキスト分析や CodeMender のコード修復と連携し、手動対応からマシンスピードの防御へシフトすることで、攻撃経路の遮断と継続的な脅威ハンティングを可能にする。Claude Security や Codex Security がシステム開発工程におけるセキュリティ強化に主な焦点を置くのに対し、本ソリューションはセキュリティ運用プロセス全体を広く支援する点に大きな違いがある。

情報源

- "Detecting and containing AI-powered threats with Google Security Operations agents | Google Cloud Blog" (2026-06-10) https://cloud.google.com/blog/products/identity-security/detecting-and-containing-powered-threats-with-google-security-operations-agents?hl=en
- "Google AI Threat Defense 発表：攻撃者の先を行くために | Google Cloud 公式ブログ" (2026-05-28) https://cloud.google.com/blog/ja/products/identity-security/introducing-google-ai-threat-defense

2.12. [2026-06-02] Fragnesia 等の Linux カーネルにおける一連の脆弱性

Fragnesia（CVE-2026-46300）を含む Linux カーネルの一連の脆弱性が確認された。これらは Copy-on-Write ページキャッシュの根本的な欠陥を悪用し、XFRM ESP-in-TCP サブシステムを通じてページキャッシュ内の読取専用ファイルを改ざんすることで権限昇格を行う。個別の CVE は異なるが、悪用の手法は共通している。検証によれば、OpenShift のような多層防御アーキテクチャ（SELinux や名前空間の制限）は、カーネルが脆弱な状態でも攻撃を効果的に阻止可能である。組織は個別のパッチ対応だけでなく、挙動監視と中央集権的なポリシー管理を重視すべきである。なお、Fragnesia は AI エージェント型のソフトウェア監査ツール V12 によって発見された。

情報源

- "Fragnesia and friends: When page cache vulnerabilities keep coming back" (2026-06-02)
  https://www.redhat.com/en/blog/fragnesia-and-friends-when-page-cache-vulnerabilities-keep-coming-back
- "pocs/fragnesia at main · v12-security/pocs · GitHub"
  https://github.com/v12-security/pocs/tree/main/fragnesia

2.13. [2026-06-02] X.Org Server 等における 9 件のメモリ安全性脆弱性

X.Org Server および XWayland において、スタックバッファオーバーフロー、Use-After-Free、境界外アクセスを含む 9 件のメモリ安全性脆弱性が開示された。これらは 8 件が TrendAI の FENRIR 静的解析ツールによって発見され、レガシーな C 言語コードベースに対する AI 駆動の監査の有効性を示した。対象は xorg-server 21.1.23 および xwayland 24.1.12 以前のバージョンである。悪用された場合、特に X サーバーが特権で動作している環境では、情報漏洩やローカル特権昇格につながる恐れがある。XWayland 経由で広く利用されているため、管理者は直ちにパッケージを更新し、リスクを軽減する必要がある。

情報源

- "X.Org Server Fixes Nine Flaws: AI Found Eight in Fourth Batch This Year" (2026-

06-04）
https://www.techtimes.com/articles/317764/20260604/xorg-server-fixes-nine-flaws-ai-found-eight-fourth-batch-this-year.htm

- "X.Org Security Advisory: multiple security issues X.Org X server and Xwayland"（2026-06-02）
https://lists.x.org/archives/xorg-announce/2026-June/003702.html
- "X.Org Server Starts June With Nine New Security Vulnerabilities Discovered Via AI - Phoronix"（2026-06-01）
https://www.phoronix.com/news/X.Org-9-Vulnerabilities-AI

2.14. [2026-06-02] AI 音声クローン詐欺対策としての「偽通話検知機能」

Google は、AI による音声クローンや発信者番号のなりすましという脅威に対抗するため、Android 向けに「偽通話検知機能」を導入した。巧妙なディープフェイク攻撃による世界的経済損失が増大する中、本機能はエンドツーエンドで暗号化された RCS（Rich Communication Services）を活用し、デバイス間で「デジタル・ハンドシェイク」を確立する。通話受信時、システムはリアルタイムで送信元デバイスと信号を照合し、期待される通信確認が取れない場合はユーザーに警告を発する。オープンスタンダードである RCS を採用することで、エコシステム全体での拡張可能な防御基盤の構築と、巧妙ななりすまし詐欺のリスク低減を目指している。

情報源

- "Android introduces fake call detection to stop deepfake scams"（2026-06-02）
https://blog.google/security/android-fake-call-detection/

2.15. [2026-06-03] HTTP/2 プロトコルにおける新規 DoS 攻撃「HTTP/2 Bomb」

主要な Web サーバー（Nginx、Apache httpd、IIS、Envoy、Cloudflare Pingora）に影響を与える新たなリモート DoS 攻撃「HTTP/2 Bomb」が判明した。本攻撃は、HPACK 圧縮爆弾と HTTP/2 のフロー制御停止を組み合わせることでメモリ枯渇を引き起こす。攻撃者は HPACK 動的テーブルとゼロバイトウィンドウ更新を悪用し、極

めて低い帯域幅で数 GB のメモリを占有できる。この攻撃は RFC 7541 およびサーバー側のメモリ管理の不備を突いており、従来のデコード済みサイズ制限では防げない。対策として、修正済みバージョンへの更新や HTTP/2 の無効化、ヘッダー数の厳格な制限、およびプロセスごとのメモリ制限が推奨される。なお、この攻撃手法は Codex が発見した。

情報源

- "Codex Discovered a Hidden HTTP/2 Bomb - Calif" (2026-06-03)
  https://blog.calif.io/p/codex-discovered-a-hidden-http2-bomb

3. AI を悪用したサイバー攻撃への対処

3.1. [2026-03-02] Trivy エコシステムを狙った継続的なサプライチェーン攻撃

2026 年 3 月、Trivy エコシステムは長期的なサプライチェーン攻撃を受けた。攻撃者は GitHub リポジトリの認証情報を侵害し、悪意のあるバイナリ（v0.69.4）、コンテナイメージ、GitHub Actions（trivy-action、setup-trivy）を配布した。攻撃手法にはタグハイジャックや悪意あるコミットが含まれ、環境変数を窃取する情報窃取型マルウェアが仕込まれた。さらに、OpenVSX 上の Trivy VS Code 拡張機能を標的とした AI 支援型攻撃も発生し、ローカルの AI コーディングアシスタントを悪用してシステムデータを外部へ流出させようとした。この脅威は Checkmarx KICS、LiteLLM、Telnyx、Axios など AI システムでも利用される他のプロジェクトにも波及しており、専門家は直ちに認証情報のローテーション、GitHub Actions の SHA ピン留め、依存関係の厳格な管理を行うよう警告している。

情報源

- "Trivy ecosystem supply chain temporarily compromised · Advisory · aquasecurity/trivy · GitHub" (2026-03-21)
  https://github.com/aquasecurity/trivy/security/advisories/GHSA-69fq-xp46-6x23
- "Trivy、LiteLLM、Axios のソフトウェアサプライチェーン侵害事案を考察～CI/CD 環境のソフトウェアサプライチェーン侵害と連鎖とは？ | トレンドマイクロ（JP)" (2026-04-09)
  https://www.trendmicro.com/ja_jp/jp-security/26/d/expertview-20260409-01.html
- "Unauthorized AI Agent Execution Code Published to OpenVSX in Aqua Trivy VS Code Extension" (2026-03-02)
  https://socket.dev/blog/unauthorized-ai-agent-execution-code-published-to-openvsx-in-aqua-trivy-vs-code-extension

3.2. [2026-04-07] Anthropic「Claude Mythos Preview」他のサイバー攻撃能力

Claude Mythos Preview のリリースは AI サイバーセキュリティにおける転換点であ

り、特に脆弱性の特定において、前例のない高度なサイバー攻撃能力に最先端AIが到達していることを明らかにした。Anthropicの内部評価や英AISIによる検証では、Mythosが主要OSやブラウザにおけるゼロデイ脆弱性を発見したのに加え、かつてはAIには不可能と考えられていた多段ROPチェーンなどの高度なエクスプロイト構築ができることが示された。しかし、AISLEやVidocによる競合分析の結果、小規模なオープンウェイトモデルであっても、専門家が設計した発見用スキャフォールドに統合すれば、FreeBSDのNFSエクスプロイトやOpenBSDの27年前のSACKバグといった、Mythos評価に示された代表的な脆弱性分析をある程度再現可能であることも指摘されている。これは、フロンティアモデルだけが高度な能力を独占的に有しているのではないことを示している。つまり、能力のフロンティアは「ギザギザ」であり、タスクごとに最適なモデルが異なることを意味する。さらに、GPT-5.5等の他の高性能モデルの登場により、能力が急速に向上するトレンドは業界全体のものとなっている。また、実際のサイバー攻撃能力は、エージェント・スキャフォールド、ツール連携、トリアージ・パイプラインといった「ハーネス」に大きく依存している。Vidocの報告では、かつては人間の専門家が数週間要したパッチ差分解析やリバースエンジニアリングをAIで自動化できるようになったことを踏まえ、N-dayエクスプロイト開発の障壁が崩壊していることを強調している。これらの進歩にもかかわらず、発見された脆弱性は従来の分類に収まるものと見られ、防御戦略の核心は不変である。メモリ安全な言語への移行、パッチ適用の迅速化、脆弱性管理の自動化といった確立されたセキュリティ習慣の徹底が依然として不可欠である。このことは、英AISIのサイバーレンジ評価では、防御措置がない状態でのMythosの攻撃成功率がさほど高くないという結果からも補強される。AIの進歩は、自律的な攻撃能力のコモディティ化をもたらしており、セキュリティ側でも、手動解析・対応から自動化された高速応答システムへと移行することを迫っていると言える。

情報源

- "AI Cybersecurity After Mythos: The Jagged Frontier | AISLE" (2026-04-07) https://aisle.com/blog/ai-cybersecurity-after-mythos-the-jagged-frontier
- "Assessing Claude Mythos Preview's cybersecurity capabilities" (2026-04-07) https://red.anthropic.com/2026/mythos-preview/
- "CAISI Evaluation of DeepSeek V4 Pro" (2026-05-01)

https://www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro

- "Measuring LLMs' ability to develop exploits" (2026-05-22)
  https://red.anthropic.com/2026/exploit-evals/
- "How fast is autonomous AI cyber capability advancing?" (2026-05-13)
  https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing
- "Measuring LLMs' impact on N-day exploits" (2026-06-08)
  https://red.anthropic.com/2026/n-days/
- "Our evaluation of OpenAI's GPT-5.5 cyber capabilities" (2026-04-30)
  https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities
- "Our evaluation of Claude Mythos Preview's cyber capabilities" (2026-04-13)
  https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities
- "We Reproduced Anthropic's Mythos Findings With Public Models" (2026-04-14)
  https://blog.vidocsecurity.com/blog/we-reproduced-anthropics-mythos-findings-with-public-models

3.3. [2026-04-07] Anthropic「Project Glasswing」

Anthropic の新しい AI モデル「Claude Mythos Preview」は、主要 OS やブラウザにおける複雑なゼロデイ脆弱性を自律的に特定・悪用する前例のない能力を示した。この進展を受け、AWS、Google、Microsoft 等のテック企業が参画する防御的枠組み「Project Glasswing」が発足した。同プロジェクトは、重要インフラの防御強化に Mythos を活用することを目指している。同モデルは防御側の脆弱性発見を加速させる一方、悪用時には甚大な国家安全保障上のリスクを伴うため、アクセスは厳格に制限されている。業界全体で、AI を用いた高速度なサイバー攻撃に対抗すべく、AI による自動化されたセキュリティライフサイクルへの移行が進んでいる。

情報源

- "Anthropic 最新 AI 巡り緊急会合、財務長官と FRB 議長が米銀 CEO を招集 -

Bloomberg" (2026-04-10)
https://www.bloomberg.com/jp/news/articles/2026-04-10/TD95ABT9NJMP00#gsc.tab=0

- "Behind the Scenes Hardening Firefox with Claude Mythos Preview - Mozilla Hacks" (2026-05-07)
https://hacks.mozilla.org/2026/05/behind-the-scenes-hardening-firefox/
- "China Sought Access to Anthropic's Newest A.I. The Answer Was No. - The New York Times" (2026-05-12)
https://www.nytimes.com/2026/05/12/us/politics/china-ai-anthropic-openai-mythos-chatgpt.html
- "米国防総省が「ミュトス」導入、アンソロピック排除は継続 | ロイター" (2026-05-13)
https://jp.reuters.com/markets/global-markets/GEXVNECE4FPJFIHZ37XWOBL6GA-2026-05-13/
- "Project Glasswing: Securing critical software for the AI era ¥ Anthropic" (2026-04-07)
https://www.anthropic.com/glasswing

3.4. [2026-05-11] TanStack npm パッケージのサプライチェーン攻撃

TeamPCP グループが GitHub Actions のキャッシュ汚染と OIDC トークン抽出を連鎖させ、42 個の@tanstack/* npm パッケージを侵害した。有効な SLSA 証明書を持つ悪意あるパッケージを生成する「Mini Shai-Hulud」ワームは、クラウド環境や CI/CD パイプライン、開発者端末から認証情報を窃取した。マルウェアは、ホームディレクトリを破壊する「デッドマンズスイッチ」やエディタの永続化フックを備える。被害者は、秘密情報のローテーション前に永続化アーティファクトを確実に除去する必要がある。組織は、OIDC の粒度設定やインストール時の挙動監視による対策が不可欠である。TeamPCP が意図的に公開したと見られるマルウェアには、AI によって開発されたものだとのメッセージが付されていた。

情報源

- "Compromised Nx Console version 18.95.0 · Advisory · nrwl/nx-console" (2026-05-18)
  https://github.com/nrwl/nx-console/security/advisories/GHSA-c9j4-9m59-847w
- "Malware crew TeamPCP open-sources its Shai-Hulud worm on GitHub" (2026-05-13)
  https://www.theregister.com/security/2026/05/13/malware-crew-teampcp-open-sources-its-shai-hulud-worm-on-github/5239319
- "Malware in 42 @tanstack/* packages exfiltrates cloud credentials, GitHub tokens, and SSH keys" (2026-05-11)
  https://github.com/TanStack/router/security/advisories/GHSA-g7cv-rxg3-hmpx
- "TanStack Npm Packages Compromised Inside The Mini Shai Hulud Supply Chain Attack" (2026-05-11)
  https://snyk.io/jp/blog/tanstack-npm-packages-compromised/
- "TeamPCP's Mini Shai-Hulud Is Back: A Self-Spreading Supply Chain Attack Compromises TanStack npm Packages" (2026-05-11)
  https://www.stepsecurity.io/blog/mini-shai-hulud-is-back-a-self-spreading-supply-chain-attack-hits-the-npm-ecosystem
- "Npm registry sets stage for more secure package publishing" (2026-05-21)
  https://www.theregister.com/ai-ml/2026/05/21/npm-registry-sets-stage-for-more-secure-package-publishing/5244527

3.5. [2026-05-18] Nx Console 経由での GitHub 内部リポジトリ侵害

脅威グループ TeamPCP（UNC6780）により約 3,800 件の GitHub 内部リポジトリが侵害された。攻撃の発端は、VS Code 拡張機能「Nx Console」（v18.95.0）の不正版によるサプライチェーン攻撃であり、18 分間だけ露出したこの不正版に感染したことで、GitHub 従業員のワークステーションから認証情報が窃取された。この侵害により、攻撃者は社内ソースコードや内部データを流出させた。GitHub は対応として、機密情報と GitHub Enterprise Server（GHES）の署名鍵をローテーションし、エンタープライズ顧客に手動での鍵更新を要請した。本件は、開発ツールに潜むリスクと、

CI/CDパイプラインの強化およびVS Code拡張機能の厳格な管理の重要性を浮き彫りにした。

情報源

- "Compromised Nx Console version 18.95.0 · Advisory · nrwl/nx-console · GitHub" (2026-05-18)
  https://github.com/nrwl/nx-console/security/advisories/GHSA-c9j4-9m59-847w
- "GitHub Confirms Hack Impacting 3,800 Internal Repositories - SecurityWeek" (2026-05-20)
  https://www.securityweek.com/github-confirms-hack-impacting-3800-internal-repositories/
- "GitHub Source Code Breach - TeamPCP Claims Access to Internal Source Code" (2026-05-20)
  https://cybersecuritynews.com/github-source-code-breach/
- "Investigation update: GitHub Enterprise Server signing key rotation - The GitHub Blog" (2026-05-26)
  https://github.blog/security/investigating-unauthorized-access-to-githubs-internal-repositories/

3.6. [2026-05-12] Google Threat Intelligence GroupによるAI悪用脅威トレンド分析

Google脅威インテリジェンスグループ（GTIG）は、脅威アクターによる生成AI活用の高度化を報告した。攻撃者はAIをゼロデイ脆弱性の発見やポリモーフィック型マルウェアの開発、PROMPTSPYに見られるような自律的な攻撃オーケストレーションに利用している。また、制限を回避するためにミドルウェア経由でAIを産業規模で利用する傾向が強まっている。同時に、オープンソースライブラリや統合コンポーネントといったAIサプライチェーンが初期アクセスの標的となっており、UNC6780のようなアクターによる情報窃取やランサムウェア攻撃が確認されている。AIは攻撃の強力な加速装置であると同時に、極めて重要なセキュリティターゲットとなっている。これらの攻撃はAnthropic「Claude Mythos Preview」が世界的に話題になる前に展開さ

れたものであり、AI の能力向上がサイバー脅威を助長する傾向が全体的なものであることを示唆している。

情報源

- "GTIG AI 脅威トラッカー: 攻撃者による脆弱性悪用、オペレーションの強化、初期アクセスのための AI 活用 | Google Cloud 公式ブログ" (2026-05-12) https://cloud.google.com/blog/ja/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access

3.7. [2026-05-18] 国家サイバー統括室による「Project YATA-Shield」発表

「Project YATA-Shield」は、フロンティア AI モデルの進化と悪用に伴うサイバー脅威に対抗するための、我が国の包括的な国家戦略パッケージである。Claude Mythos や GPT-5.5-Cyber 等の高性能 AI が脆弱性の発見から悪用までの期間を「月単位から時間単位」へ劇的に短縮する現実を踏まえ、本パッケージは関係各所に対策を要求している。まず、重要インフラ事業者と政府機関には、経営層・組織トップの主導のもと、ゼロトラスト設計への移行、厳格なパッチ管理、およびシステム侵害を前提とした「隔離・復旧」体制の整備を求めている。次に、ソフトウェアベンダには「セキュア・バイ・デザイン」の原則に基づき、開発段階から高性能 AI を活用した脆弱性の早期発見・是正に努めるよう注意喚起する。日本の AI セーフティ・インスティテュート（AISI）においては、技術的支援、AI モデルの安全性評価や国際連携を推進する。

情報源

- “「AI 性能の高度化を踏まえたサイバーセキュリティ対策に関する関係省庁会議」を開催” (2026-05-18) https://www.cyber.go.jp/news/list/index.html

3.8. [2026-05-19] AI コンテンツ来歴管理に向けた OpenAI と Google の連携強化

OpenAI と Google は、AI コンテンツの透明性を向上させる取り組みとして、C2PA 標準と SynthID ウォーターマークの統合を発表した。OpenAI は C2PA 準拠を達成し、

プラットフォーム間でメタデータの耐久性を確保するSynthIDを導入、公開検証ツールも提供した。同時にGoogleは、Search、Gemini、Pixelデバイス等のエコシステム全体でSynthIDとC2PAの統合を拡大し、企業向けAIコンテンツ検出APIも発表した。これらの協調的な取り組みは、相互運用可能な来歴エコシステムを構築し、ユーザーがメディアの出所や履歴を検証可能にすることで、偽造合成コンテンツのリスクを低減し、多層的な検出手法を通じてデジタルメディアの信頼性を高めることを目指している。

情報源

- "より安全で透明性の高い AI エコシステムに向けて、コンテンツ来歴の取り組みを前進 | OpenAI" (2026-05-19)
  https://openai.com/ja-JP/index/advancing-content-provenance/
- "Tools to understand how content was created and edited" (2026-05-19)
  https://blog.google/innovation-and-ai/products/identifying-ai-generated-media-online/

3.9. [2026-05-21] サイバー防御特化モデル OpenAI「GPT-5.5-Cyber」提供開始

OpenAIは重要インフラ防衛を支援する特化型AIモデル「GPT-5.5-Cyber」を発表した。審査を経た特定組織等にアクセスを提供する「Trusted Access for Cyber (TAC)」フレームワークを通じ、脆弱性トリアージやマルウェア解析、リバースエンジニアリング等の専門的ワークフローにおけるセーフガードを最適化する。利用には厳格な本人確認とアカウントセキュリティが必須となる。日本政府とも導入に向けた協議を開始しており、サイバー防御能力の向上と悪用防止のバランスを両立させる方針である。

情報源

- "オープンＡＩが最新モデルを日本の政府・企業に提供へ…取締役が記者会見、「クロード・ミュトス」匹敵の性能 ： 読売新聞" (2026-05-21)
  https://www.yomiuri.co.jp/economy/20260521-GYT1T00240/

- "GPT-5.5 と GPT-5.5-Cyber で Trusted Access for Cyber を拡大 | OpenAI" (2026-05-07)
  https://openai.com/ja-JP/index/gpt-5-5-with-trusted-access-for-cyber/

3.10. [2026-06-02] AI エージェントを活用した適応型コンピュータワーム

CleverHans Lab の研究チームは、自律的な推論によりネットワークの脆弱性を特定・悪用する AI 駆動型の自己複製ワームを発表した。従来の静的なマルウェアとは異なり、このエージェントはローカルのオープンウェイト LLM を使用して攻撃戦略をリアルタイムで適応させる。被害者の計算リソースを寄生的に利用することで攻撃コストをほぼゼロに抑え、中央集権的なベンダーの安全フィルターを回避する。この概念実証は、再帰的推論とツール利用を組み合わせることで、自律的なサイバー攻撃が現実の脅威となることを示している。このような適応力の高い攻撃者に対抗するためには、ゼロトラストアーキテクチャやマイクロセグメンテーション、迅速な自動パッチ適用への移行が不可欠である。

情報源

- "CleverHans Lab - AI Agents Enable Adaptive Computer Worms" (2026-06)
  https://cleverhans.io/worm.html

3.11. [2026-06-03] LLM ATT&CK Navigator による AI 駆動型サイバー脅威分析

Anthropic の報告書「LLM ATT&CK Navigator」は、2025 年 3 月から 2026 年 3 月までの AI 悪用アカウント 832 件を分析した。攻撃者は当初、主にマルウェア開発などの準備段階に AI を利用していたが、現在はラテラルムーブメントや認証情報窃取といった「侵害後」の活動へシフトしている。1 年間で中リスク以上の攻撃者の割合は 33% から 56%へ急増しており、これは AI エージェントによる攻撃ステージの自律的な連鎖（スキャフォールディング）が要因である。既存の MITRE ATT&CK フレームワークではこれらの自律的な挙動を十分に捉えきれておらず、今後は技術スキルや手法数ではなく、オーケストレーションに基づくリスク評価への転換が求められるとしている。

情報源

- "LLM ATT&CK Navigator | red.anthropic.com" (2026-06-03)
  https://red.anthropic.com/2026/attack-navigator/
- "What we learned mapping a year’s worth of AI-enabled cyber threats | Anthropic" (2026-06-03)
  https://www.anthropic.com/news/AI-enabled-cyber-threats-mitre-attack