AIセキュリティ短信 2026年3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

1. AIに係る安全性確保（Security for AI）

1.1. [2025-05] Anthropic Claude Coworkでの間接プロンプトインジェクション

AIエージェントのプレビュー版であるClaude Coworkには、間接プロンプトインジェクションによるファイル流出の脆弱性が存在する。攻撃者は文書（.docx等）に隠された悪意のあるプロンプトを埋め込み、エージェントを操作して無許可のAPI呼び出しを実行させることが可能である。エージェントの仮想環境はAnthropicのAPIへの通信を許可しているため、攻撃者は埋め込まれたAPIキーを使用して、ローカルファイルを自身の環境へ送信できる。この脆弱性は、エージェントの自律的な実行能力と、データ処理における人間による検証の欠如を悪用するものである。モデルの防御力に関わらず、本質的な問題は実行環境の境界の不備と、エージェントが信頼できないコンテンツを処理する際のリスクにある。

情報源

- "Claude Cowork Exfiltrates Files"
  https://www.promptarmor.com/resources/claude-cowork-exfiltrates-files

1.2. [2025-12] ブラウザ拡張機能を悪用したAI会話データ窃取インシデント

最近の調査により、Googleの「Featured（おすすめ）」バッジが付与されたものを含む複数の悪意あるChrome/Edge拡張機能が、機密性の高いAI会話を窃取していることが判明した。攻撃者はスクリプトインジェクションを用いてブラウザAPI（fetch/XMLHttpRequest）を乗っ取るか、ChatGPT、Claude、DeepSeek等のプラットフォーム上でDOM要素をスクレイピングする手法をとっている。これらのキャンペーンにより合計890万人以上のユーザーが影響を受けた可能性がある。窃取されたプロンプトや回答、セッションメタデータは攻撃者のC2サーバーへ送信されており、企業秘密の漏洩や個人情報流出のリスクが懸念される。ストアの信頼性を悪用した配布手法が拡大を助長した。

情報源

- "8 Million Users' AI Conversations Sold for Profit by "Privacy" Extensions" (2025-12-15)
  https://www.koi.ai/blog/urban-vpn-browser-extension-ai-conversations-data-collection
- "900K Users Compromised: Chrome Extensions Steal ChatGPT and DeepSeek Conversations" (2025-12-30)
  https://www.ox.security/blog/malicious-chrome-extensions-steal-chatgpt-deepseek-conversations/

1.3. [2026-01-12] Superhuman AI における間接プロンプトインジェクション

セキュリティ研究者が Superhuman AI において、ユーザーの操作なしに機密メールを流出させる間接プロンプトインジェクションの脆弱性を特定した。攻撃者はメールや Web コンテンツに悪意ある指示を埋め込み、AI を操作して Google フォームの事前入力リンクや直接的なサーバーリクエストを通じてインボックスのデータを外部送信させた。この脆弱性は、許可されたドメインの悪用や Markdown 画像の自動レンダリング機能を通じて Content Security Policy を回避していた。Superhuman のセキュリティチームは迅速に対応し、Superhuman Go を含む影響を受ける製品群に修正を適用した。

情報源

- "Superhuman AI Exfiltrates Emails" (2026-01-12)
  https://www.promptarmor.com/resources/superhuman-ai-exfiltrates-emails

1.4. [2026-01-14] Microsoft Copilot における「Reprompt」脆弱性

Varonis Threat Labs は、Microsoft Copilot Personal において 1 クリックでデータ流出を可能にする脆弱性「Reprompt」を発見した。本攻撃は URL の「q」パラメータを用いた P2P（Personal Prompt）インジェクションを悪用し、「ダブルリクエスト」および「チェーンリクエスト」の手法でセキュリティ保護機能を回避する。Copilot にタス

クを二度実行させ、サーバーからの動的な後続コマンドを利用することで、ユーザーが最初にリンクをクリックした後の追加操作なしに機密データを隠密に窃取可能である。Microsoft は本脆弱性を修正済みであるが、この研究は AI アシスタントにおけるクライアント監視の回避やセッションの悪用といった重大なリスクを浮き彫りにした。防御策として、外部入力を信頼せず、永続的かつ多段階の AI ガードレールを実装することが推奨される。

情報源

- "Reprompt: The Single-Click Microsoft Copilot Attack that Silently Steals Your Personal Data" (2026-01-14)
  https://www.varonis.com/blog/reprompt

1.5. [2026-01-26] AI エージェントゲートウェイ「Clawdbot」の深刻な脆弱性

オープンソースの AI エージェントゲートウェイ「Clawdbot」において、深刻な認証バイパス脆弱性が原因で、900 以上のインスタンスがインターネット上に公開されていることが判明した。この問題は、リバースプロキシ配下で展開された際に X-Forwarded-For ヘッダーを適切に処理できず、「ローカルホスト自動承認」が誤作動することに起因する。これにより、認証なしでリモートから Control UI へのアクセスが可能となり、API キー（Anthropic、Slack、Telegram 等）の窃取、数ヶ月分の会話履歴の流出、さらにルート権限でのリモートコード実行を許すリスクがある。専門家は、gateway.trustedProxies の設定、パスワード認証の有効化、および侵害された可能性のある全認証情報の再発行を強く推奨している。

情報源

- "Hundreds of Exposed Clawdbot Gateways Leave API Keys and Private Chats Vulnerable" (2026-01-26)
  https://cybersecuritynews.com/clawdbot-chats-exposed/
- "hacking clawdbot and eating lobster souls" (2026-01-25)
  https://x.com/theonejvo/status/2015401219746128322

1.6. [2026-02-01] OpenClawにおけるマルウェアキャンペーン「ClawHavoc」

OpenClawのエコシステムを標的とした「ClawHavoc」キャンペーンにおいて、ClawHubマーケットプレイス上で824件の悪意あるスキルが特定された。攻撃者はタイポスクワッティングや正規に見えるドキュメントを悪用し、Atomic macOS Stealer（AMOS）を含むトロイの木馬を配布した。ペイロードは、静的解析を回避するためにパスワード保護されたアーカイブや難読化されたシェルスクリプト内に隠蔽されていた。主な攻撃手法には、偽の前提条件の要求、リバースシェル、環境ファイルからの認証情報窃取が含まれる。これを受け、OpenClawはVirusTotalと提携し、すべてのスキルに対してLLMベースの「Code Insight」を含む自動セキュリティスキャンを導入した。

情報源

- "ClawHavoc: 341 Malicious Clawed Skills Found by the Bot They Were Targeting"（2026-02-01）
  https://www.koi.ai/blog/clawhavoc-341-malicious-clawedbot-skills-found-by-the-bot-they-were-targeting
- "OpenClaw Partners with VirusTotal for Skill Security — OpenClaw Blog"（2026-02-07）
  https://openclaw.ai/blog/virustotal-partnership

1.7. [2026-02-09] AIエージェントに起因するClineへのサプライチェーン攻撃

AIコーディングツール「Cline」において、自動課題トリアージワークフローのプロンプトインジェクションに起因する重大なサプライチェーン脆弱性が確認された。攻撃者は悪意のある課題タイトルを用いてClaude-code-actionを操作し、任意のコード実行権限を獲得。さらにGitHub Actionsのキャッシュを汚染することで、VS Code MarketplaceやNPMの公開用クレデンシャル（VSCE_PAT等）を奪取した。本件は実際に悪用され、不正なCline CLIが公開される事態に発展した。AIエージェントに過度な権限を付与することが、CI/CD環境における深刻なリスクとなることを示している。

情報源

- "Clinejection — Compromising Cline's Production Releases just by Prompting an Issue Triager" (2026-02-09)
  https://adnanthekhan.com/posts/clinejection/#pre-publication

1.8. [2026-02-10] OpenClaw における設定不備と脆弱性による大量露呈

SecurityScorecard の STRIKE チームは、サービスが全インターフェースにバインドされるという不適切なデフォルト設定により、15,200 件の OpenClaw（旧 Moltbot）コントロールパネルがインターネット上に露呈していることを確認した。CVE-2026-25253（RCE）、CVE-2026-25157（SSH コマンドインジェクション）、CVE-2026-24763（Docker サンドボックスエスケープ）という高リスクの脆弱性に対し、パッチ未適用のインスタンスが多く残存している点が極めて危険である。AI エージェントはクラウド認証情報や SSH 鍵、ブラウザセッションへのアクセス権を持つため、侵害されるとホストの完全な制御権を奪取され、ラテラルムーブメントを許すことになる。特定されたインスタンスの約 33.8%には APT グループを含む既存の悪意ある活動との関連が見られる。対策として、バージョン 2026.2.1 以降への更新、localhost（127.0.0.1）へのバインド設定、およびすべての API キーのローテーションが不可欠である。

情報源

- "15,200 OpenClaw Control Panels with Full System Access Exposed to Internet" (2026-02-10)
  https://cybersecuritynews.com/openclaw-control-panels-exposed/

1.9. [2026-02-17] Boundary Point Jailbreaking によるブラックボックス攻撃

Boundary Point Jailbreaking (BPJ) は、Constitutional Classifiers や GPT-5 の入力フィルタなど、堅牢な LLM セーフガードを標的とする新規の完全自動化ブラックボックス攻撃手法である。勾配や確信度スコアを必要とする従来の手法とは異なり、BPJ は「フラグの有無」というバイナリフィードバックのみを利用する。BPJ はカリキュラム学習と、微小な変化に敏感なターゲットである「境界点（boundary points）」を活用

して敵対的プレフィックスを最適化する。BPJは、数千時間の人間によるレッドチーミングに耐えた最先端の防御に対し、普遍的な脱獄を成功させた。研究チームは、BPJが大量のクエリを生成することから、単一の対話に基づく防御では不十分であると指摘する。個別のプロンプト分類に頼るのではなく、複数のリクエストにわたる不審なパターンを検出する「バッチレベル監視」を含む多層的な防御アプローチを提唱している。

情報源

- "Boundary Point Jailbreaking of Black-Box LLMs" (2026-02-17) https://arxiv.org/abs/2602.15001v2
- "Boundary Point Jailbreaking: A new way to break the strongest AI defences | AISI Work" (2026-02-17) https://www.aisi.gov.uk/blog/boundary-point-jailbreaking-a-new-way-to-break-the-strongest-ai-defences

1.10. [2026-02-18] AIエージェントの自律性に関する実態調査と分析

Anthropicは Claude Code およびパブリック API を通じて数百万件の人とエージェントの対話を分析し、実社会における AI 自律性の実態を調査した。分析の結果、モデルの潜在能力は実際の運用レベルを上回っており、ユーザーは習熟度が増すにつれて自動承認率を高める傾向にあることが判明した。一方で、介入（中断）による監視も継続されている。特に複雑なタスクにおいて Claude Code が自発的に確認を求める動きが見られ、エージェントによる自己監視が重要な安全策となり得ることが示唆された。現状、多くの運用は低リスクだが、サイバーセキュリティや金融などへの拡大が進んでいる。結論として、画一的な承認プロセスの強制よりも、ポストデプロイメントのモニタリング強化と、アクティブな人間による監視を軸とした設計が推奨される。

情報源

- "Measuring AI agent autonomy in practice" (2026-02-18) https://www.anthropic.com/research/measuring-agent-autonomy

1.11. [2026-02-18] Microsoft 365 Copilot における機密メールの不正要約事象

Microsoft 365 Copilot（追跡 ID：CW1226324）の不具合により、AI アシスタントが「送信済みアイテム」や「下書き」フォルダ内の機密メールを不適切に要約し、設定済みのデータ損失防止（DLP）ポリシーや機密ラベルを回避する事象が発生した。2026 年 1 月 21 日に検知されたこの問題は、Copilot の「作業」タブが構成設定に反して保護されたコンテンツを読み取ってしまうコードエラーに起因する。Microsoft は、ユーザーが閲覧権限を持つ情報のみが処理されたと説明しているが、本件は AI のコンテンツ認識ガードレールにおける重大な機能不全を露呈した。現在は、Copilot が機密ラベルと DLP 制約を適切に尊重するよう、全世界で修正設定が展開されている。

情報源

- "Microsoft says bug causes Copilot to summarize confidential emails" (2026-02-18) https://www.bleepingcomputer.com/news/microsoft/microsoft-says-bug-causes-copilot-to-summarize-confidential-emails/

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

2.1. [2025-12-15] 自律型AIペネトレーションテストツール「Shannon」

Shannonは、Keygraphが開発したWebアプリケーションおよびAPI向けの自律型ホワイトボックスAIペネトレーションテストフレームワークである。ソースコード解析と実際の攻撃実行を組み合わせ、インジェクション、XSS、SSRF、認証バイパスなどの脆弱性を特定する。従来の静的解析ツールとは異なり、実際にエクスプロイトを実行して再現可能なPoCを提示することで、誤検知を最小限に抑える。マルチエージェントアーキテクチャによりレッドチームの戦術を模倣し、XBOWセキュリティベンチマークでは96.15%の成功率を記録して手動テストを凌駕した。AI導入による開発加速で生じるセキュリティギャップを埋めるため、非本番環境での継続的なオンデマンドテストを実現する。

情報源

- "GitHub - KeygraphHQ/shannon: Shannon Lite is an autonomous, white-box AI pentester"
  https://github.com/KeygraphHQ/shannon
- "Shannon - AI Pentesting Tool that Autonomously Checks for Code Vulnerabilities and Executes Real Exploits" (2025-12-15)
  https://cybersecuritynews.com/shannon-ai-pentesting-tool/

2.2. [2025-12-22] ChatGPT Atlasにおける自動レッドチームの活用

OpenAIは、ブラウザーエージェント「ChatGPT Atlas」におけるプロンプトインジェクションのリスク低減に向け、セキュリティ対策を強化した。同社は、強化学習（RL）を用いたLLMベースの自動レッドチームシステムを導入し、現実世界で悪用される前に多段階の攻撃ベクトルを特定・修正している。このシステムは、防御側モデルの推論過程への特権的なアクセスを活用することで、より精度の高い攻撃シミュレーションと迅速なパッチ適用を実現する。プロンプトインジェクションはエージェントのWebアクセス機能に起因する長期的課題であるが、自動化された攻撃発見と敵対的学習を組み合わせることで、攻撃者のコストを増大させ、リスクを実質的に低減することを目指している。

情報源

- "ChatGPT Atlas をプロンプトインジェクション攻撃に対して継続的に強化しています | OpenAI" (2025-12-22)
  https://openai.com/ja-JP/index/hardening-atlas-against-prompt-injection/

2.3. [2026-01-30] Constitutional Classifiers++によるジェイルブレイク対策

Anthropic は、LLM におけるユニバーサル・ジェイルブレイクに対する高度な防御フレームワーク「Constitutional Classifiers++」を発表した。従来システムの課題であった再構成攻撃や出力難読化攻撃に対処するため、二段階のカスケード構成を採用。軽量な内部線形活性化プローブでトラフィックをスクリーニングし、疑わしい対話のみを文脈認識型のアンサンブル分類器へエスカレーションする。これにより、計算コストを 40 倍削減し、誤拒否率を 0.05%まで低減しつつ高い堅牢性を実現した。1,700 時間以上のレッドチーミングにおいて、ユニバーサル・ジェイルブレイクの成功は報告されておらず、本番環境における実用性とセキュリティ性能の両面で大きな進化を遂げている。

情報源

- "Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming" (2025-01-30)
  https://arxiv.org/abs/2501.18837v1
- "CONSTITUTIONAL CLASSIFIERS++: EFFICIENT PRODUCTION-GRADE DEFENSES AGAINST UNIVERSAL JAILBREAKS" (2026-01-09)
  https://arxiv.org/abs/2601.04603v1
- "Next-generation Constitutional Classifiers: More efficient protection against universal jailbreaks" (2026-01-09)
  https://www.anthropic.com/research/next-generation-constitutional-classifiers

2.4. [2026-02-05] LLM を活用したゼロデイ脆弱性発見能力の向上

Anthropic の Claude Opus 4.6 は、特殊な足場を組むことなくオープンソースプロジェ

クトで500件以上の深刻なゼロデイ脆弱性を発見し、自動脆弱性探索における大きな進歩を示した。ランダムな入力を生成する従来のファザーとは異なり、Opus 4.6はGitコミット履歴の分析、危険なコードパターンの特定、複雑なロジック（GIF処理におけるLZW圧縮など）の理解といった人間のような推論を行う。この能力は防御的なパッチ適用を加速させる一方、デュアルユースのリスクも伴う。Anthropicは、モデルベースの「プローブ」を実装し、サイバー攻撃への悪用をリアルタイムで検知・防止する体制を整えた。今回の結果は、LLMが脆弱性発見において人間を凌駕する転換点に達しつつあることを示唆しており、既存の開示基準や脆弱性管理ワークフローの再検討が必要である。

情報源

- "Evaluating and mitigating the growing risk of LLM-discovered 0-days" (2026-02-05)
  https://red.anthropic.com/2026/zero-days/

2.5. [2026-02-19] Google Play と Android エコシステムのセキュリティ強化

2025年、GoogleはAI駆動型の防御と開発者向けの先制的な取り組みを通じて、Androidエコシステムのセキュリティ体制を大幅に強化した。主な成果として、175万件のポリシー違反アプリの公開阻止と、8万を超える悪質な開発者アカウントの排除が挙げられる。Google Play Protectは不正検知機能を185の市場へ拡大し、1日あたり3500億個のアプリをスキャンして2700万個ものストア外からの脅威を無力化した。さらに、通話中の詐欺対策、Play Integrity APIにおけるハードウェアベースの信号導入、Android Studioへのセキュリティチェック統合などが実装された。これらの施策は、過度なデータアクセスの抑制、サイドローディングアプリによる金融詐欺の防止、および身元確認による開発者の説明責任強化に焦点を当てている。

情報源

- "Keeping Google Play & Android app ecosystems safe in 2025" (2026-02-19)
  https://security.googleblog.com/2026/02/keeping-google-play-android-app-ecosystem-safe-2025.html

2.6. [2026-02-20] Claude Code Security による高度な脆弱性検出・修正の支援

Anthropic は、従来の静的解析ツールでは見落とされがちな複雑な脆弱性を特定・修正するための AI 駆動型ツール「Claude Code Security」の限定プレビュー版を公開した。ルールベースのスキャナーとは異なり、本プラットフォームは Claude の推論能力を活用してビジネスロジックやデータフローを分析し、人間のようなセキュリティ研究者の挙動を模倣する。誤検知を最小限に抑える多段階検証プロセスを備え、人間のレビュー用に修正パッチを提案する。Claude Opus 4.6 の能力を活用しており、同モデルは最近オープンソースプロジェクトで 500 件以上の長期間放置された脆弱性を発見した実績がある。本ツールは、コードベースを能動的に保護することで、AI を活用したサイバー脅威から防御者を支援することを目的としている。現在はエンタープライズ、チーム、およびオープンソースのメンテナ向けに限定公開されている。

情報源

- "Making frontier cybersecurity capabilities available to defenders" (2026-02-20) https://www.anthropic.com/news/claude-code-security

2.7. [2026-02-25] Android における AI 駆動型モバイル詐欺対策の進化と普及

Google は、高度化するモバイル詐欺に対抗するため、オンデバイス AI および Gemini モデルを活用したセキュリティ戦略を強化している。リアルタイムの通話スクリーニングやメッセージの会話分析を含む Android の多層防御機能は、ソーシャルエンジニアリングや「ピッグ・ブッチャリング（投資詐欺）」、フィッシング攻撃の抑制を目的としている。Counterpoint Research や Leviathan Security Group の評価では、Android が iOS よりも AI 駆動型セキュリティ機能で先行していることが示された。Google は、これまで Pixel 限定であったこれらのプロアクティブな保護機能を、Samsung Galaxy S26 シリーズなど他の Android デバイスへ拡大しており、通信データのオンデバイス処理によるプライバシー保護を強調している。

情報源

- "Assessing the State of AI-Powered Mobile Security" (2025-10-30)
  https://counterpointresearch.com/en/insights/assessing-the-state-of-ai-powered-mobile-security-blog
- "How Android provides the most effective protection to keep you safe from mobile scams" (2025-10-30)
  https://security.googleblog.com/2025/10/how-android-protects-you-from-scams.html
- "Staying One Step Ahead: Strengthening Android's Lead in Scam Protection" (2026-02-25)
  https://security.googleblog.com/2026/02/strengthening-android-lead-in-scam-protection.html

2.8. [2026-03-06] AIを活用したMozilla Firefoxの脆弱性特定と修正

Anthropicの Frontier Red Teamは、Claude Opus 4.6を用いてFirefoxのコードベースから14件の重大な脆弱性を含む計22件の脆弱性を特定した。この協業により、AI支援型分析がMozillaのセキュリティワークフローに統合され、Firefox 148での迅速な修正につながった。本研究は、AIエージェントがコードとパッチを自己検証する「タスク・ベリファイア」が、従来のファジングでは見逃されがちな論理エラーの特定に有効であることを実証した。Claudeは脆弱性発見において高い能力を示した一方、悪用（エクスプロイト）開発能力は限定的であったが、両者のギャップは縮まりつつある。開発者は、新たな脅威に対する防御を強化するため、AIを活用したセキュリティツールの導入を急ぐべきである。

情報源

- "Hardening Firefox with Anthropic's Red Team" (2026-03-06)
  https://blog.mozilla.org/en/firefox/hardening-firefox-anthropic-red-team/
- "Partnering with Mozilla to improve Firefox's security" (2026-03-06)
  https://www.anthropic.com/news/mozilla-firefox-security

2.9. [2026-03-06] OpenAI「Codex Security」の発表

OpenAIは、プロジェクトの文脈や構造を分析してアプリケーションセキュリティを強化するエージェントツール「Codex Security」（旧称：Aardvark）を発表した。フロンティアモデルを活用し、編集可能な脅威モデルの生成、サンドボックス環境での脆弱性検証、修正パッチの提案を行う。ベータテストではアラートの84%削減、誤検知率の50%以上の低下を達成している。すでにOpenSSH、GnuTLS、GOGSなどの主要なOSSプロジェクトで深刻な脆弱性を特定しており、今後はChatGPT Enterprise等のユーザーへの提供に加え、オープンソースメンテナー支援プログラムも拡充する。

情報源

- "Codex Security が研究プレビュー版として利用可能に | OpenAI" (2026-03-06) https://openai.com/ja-JP/index/codex-security-now-in-research-preview/

2.10. [2026-03-09] Anthropic「Code Review」の発表

Anthropicは、GitHubと統合されたプルリクエスト解析ツール「Code Review」を発表した。軽量なリンターとは異なり、複数のAIエージェントが並列でコードを詳細に解析し、論理エラー、脆弱性、回帰リスクを特定する。誤検知を減らす検証ステップを備え、重要度別にインラインコメントを投稿する。組織はCLAUDE.mdやREVIEW.mdを使用してレビュー規約をカスタマイズ可能である。現在、TeamおよびEnterpriseプラン向けにリサーチプレビューとして提供されており、利用料はトークンベース（1回平均15〜25ドル）である。本ツールは、人間が見落としがちな潜在的なバグや認証の不備を指摘することで、コードレビューのボトルネック解消を目的としている。

情報源

- "Code Review - Claude Code Docs" https://code.claude.com/docs/en/code-review
- "Bringing Code Review to Claude Code" (2026-03-09) https://claude.com/blog/code-review

3. AIを悪用したサイバー攻撃への対処

3.1. [2026-02] AIを活用した大規模なFortiGate不正アクセスキャンペーン

2026年1月〜2月、ロシア語圏の脅威アクターが生成AIサービスを悪用し、世界中の600台以上のFortiGateデバイスを侵害した。本キャンペーンはゼロデイ脆弱性ではなく、公開された管理ポートに対するスキャンと脆弱な認証情報を悪用した。AIは偵察の自動化、カスタムツールの開発、攻撃計画の生成を行う「戦力倍増ツール」として機能した。アクターは「ARXON」と呼ぶ独自のModel Context Protocol（MCP）フレームワークを用い、LLMを侵入ワークフローに統合して横展開、認証情報奪取（DCSync/Impacket）、バックアップ基盤（Veeam）への攻撃を遂行した。アクター自体の技術力は中程度だが、AIによる支援で運用規模が大幅に拡大している。対策には、管理インターフェースの公開停止、MFAの強制、バックアップシステムの堅牢化が不可欠である。

情報源

- "AI-augmented threat actor accesses FortiGate devices at scale | AWS Security Blog" (2026-02-20)
  https://aws.amazon.com/jp/blogs/security/ai-augmented-threat-actor-accesses-fortigate-devices-at-scale/
- "Amazon: AI-assisted hacker breached 600 Fortinet firewalls in 5 weeks" (2026-02-21)
  https://www.bleepingcomputer.com/news/security/amazon-ai-assisted-hacker-breached-600-fortigate-firewalls-in-5-weeks/
- "LLMs in the Kill Chain: Inside a Custom MCP Targeting FortiGate Devices Across Continents" (2026-02-21)
  https://cyberandramen.net/2026/02/21/llms-in-the-kill-chain-inside-a-custom-mcp-targeting-fortigate-devices-across-continents/

3.2. [2026-02-05] GPT-5.3-Codexのサイバーセキュリティ機能と安全対策

OpenAIのGPT-5.3-Codexは、エージェント型コーディングモデルとして初めて

「Preparedness Framework」におけるサイバーセキュリティ分野で「High（高度）」の能力を有すると評価された。エンドツーエンドのサイバー操作、脆弱性発見、エクスプロイト生成において高い能力を示す一方、OpenAI はサンドボックス環境での実行、安全性推論器を用いた会話監視、および高リスク機能を制限する「Trusted Access for Cyber (TAC)」プログラムを導入した。本モデルは先行モデルと比較して、長期的な自律性と運用の一貫性が大幅に向上している。OpenAI は、依然として存在するジェイルブレイクのリスクや技術のデュアルユース性を認識しつつ、サイバーセキュリティエコシステムを支援するための防御策への投資を継続している。

情報源

- "GPT-5.3-Codex System Card - OpenAI Deployment Safety Hub" (2026-02-05) https://deploymentsafety.openai.com/gpt-5-3-codex/disallowed-content-evaluations
- "Update to GPT-5 System Card: GPT-5.2 - OpenAI Deployment Safety Hub" (2025-12-11) https://deploymentsafety.openai.com/gpt-5-2/cybersecurity

3.3. [2026-02-26] メキシコ政府機関に対する生成 AI を悪用したサイバー攻撃

脅威アクターが Anthropic の Claude を悪用し、メキシコ政府機関に対して約 1 ヶ月間にわたるサイバー攻撃を実行。150GB の機密データを流出させた。攻撃者はプロンプトエンジニアリングとプレイブック手法を用いて AI の安全ガードレールを回避し、攻撃計画の作成や横展開（ラテラルムーブメント）の指示を Claude から引き出した。Claude の制限に直面した際には OpenAI の ChatGPT を併用していた。本件は AI を活用した攻撃の急増を浮き彫りにしており、防衛側はエッジデバイス、アイデンティティ管理、クラウド/SaaS、そして AI エージェントインフラという 4 つのドメインに対する防御を強化する必要がある。従来のマルウェア手法から脱却し、ファイルレスかつ AI 主導の攻撃が増加する中、ドメイン横断的な監視とアイデンティティ中心のゼロトラストアーキテクチャの構築が不可欠である。

情報源

- "Claude didn't just plan an attack on Mexico's government. It executed one for a month — across four domains your security stack can't see. | VentureBeat" (2026-02-26)
  https://venturebeat.com/security/claude-mexico-breach-four-blind-domains-security-stack
- "Hacker Used Anthropic's Claude to Steal Sensitive Mexican Data - Bloomberg" (2026-02-25)
  https://www.bloomberg.com/news/articles/2026-02-25/hacker-used-anthropic-s-claude-to-steal-sensitive-mexican-data

3.4. [2026-03-06] 脅威アクターによるAI運用と攻撃ライフサイクルへの統合

脅威アクターは、攻撃の効率化とスケールアップを目的に、サイバー攻撃ライフサイクル全体でAIを運用化している。Microsoftの脅威インテリジェンスによると、Jasper SleetやCoral Sleetといった北朝鮮系アクターが、フィッシング、コード生成、人物の捏造、インフラ管理にLLMを活用している。主な手口には、ジェイルブレイクによる安全フィルタの回避や、反復的な意思決定のためのエージェントワークフローの活用が含まれる。人間が標的の制御権を保持しつつ、AIが偵察、マルウェア開発、侵害後の活動における技術的障壁を下げる「フォースマルチプライヤー（戦力倍増ツール）」として機能する。防御には、強力なID保護、行動分析、AI固有のセキュリティガバナンスに加え、悪意あるモデル活用の検知と緩和が必要である。

情報源

- "AI as tradecraft: How threat actors operationalize AI | Microsoft Security Blog" (2026-03-06)
  https://www.microsoft.com/en-us/security/blog/2026/03/06/ai-as-tradecraft-how-threat-actors-operationalize-ai/

3.5. [2026-03-09] McKinseyのAIシステムへのAIエージェントによる攻撃

2026年3月、CodeWallの研究チームは自律型攻撃AIエージェントを用い、McKinsey & Companyの内部AIプラットフォーム「Lilli」における重大な脆弱性を実証した。エージェントは公開されていたAPIエンドポイントに未認証のSQLインジェクションを発見し、本番データベースへの完全な読み書き権限を取得した。これにより、4,650万件のチャットメッセージや機密文書、独自研究データが漏洩。さらに、AIの「プロンプト層」への書き込み権限を獲得したことで、AIの回答操作やガードレールの無効化といった深刻なリスクが浮き彫りとなった。本件は、AIプロンプトが整合性管理やアクセス制御が不十分なまま「最重要資産」化している現状を浮き彫りにしている。

情報源

- "How We Hacked McKinsey's AI Platform — CodeWall.ai" (2026-03-09)
  https://codewall.ai/blog/how-we-hacked-mckinseys-ai-platform