NIDS 防衛研究所 National Institute for Defense Studies
NIDSコメンタリー
第 434 号　2026 年 5 月 19 日

高度なサイバー攻撃キャンペーンにおける AI 悪用によるスケール化のインパクトと対策

先進領域研究部サイバー安全保障研究室特任研究員　佐々木　勇人

はじめに

2026 年 2 月 25 日に Bloomberg などが、「メキシコ政府機関が Claude を使ったサイバー攻撃に遭い、1 億 9500 万人あまりの納税記録や住民情報が漏洩した」と報じた[1]。調査を行ったセキュリティ企業 GambitSecurity 社によると[2]、各 AI サービスには、サイバー攻撃用途の悪用などを防ぐ対策がなされているが、この攻撃者は同サービスの利用にあたり、脆弱性を探すペネトレーションテストやバグバウンティ目的であるように装い、Anthropic 社の Claude Code の悪用防止策を回避しようとしていたとされる。具体的には、Claude 側の警告などを回避する手法を駆使するほか、Claude 側に利用を拒否された場合は、ChatGPT も併用するなどして情報を得ていたとされる。被害を指摘されたうち、国家選挙管理委員会（INE）やハリスコ州政府は被害を否定しているが、現時点（本稿執筆時）では、その詳細について政府からの公式発表等はなされていない。

AI サービス自体のセキュリティ上の問題点が指摘されるとともに、サイバー攻撃者が AI サービスを悪用してその攻撃手法の高度化や活動の拡大・高速化を狙う可能性は以前から指摘されてきたところ、ここ 1〜2 年の間に具体的な悪用事例[3]、特に本稿で解説するような、APT（Advanced Persistent Threat）といった高度なサイバー攻撃活動での悪用が顕在化している状況にある。

APT などの高度なサイバー攻撃活動は、その目的達成のためにセキュリティ製品の検出を回避しながらの長期潜伏や、痕跡をなるべく残さない侵害拡大（横展開：ラテラルムーブメントと呼ばれる）など、技術的に高度なスキルを必要とするため、人的リソースやスキルのレベルにより攻撃範囲や活動頻度が制約を受けることから、ある意味、限定的な範囲・期間しか顕在化してこなかった。本稿では最近の悪用事例を紹介しつつ、AI サービスの悪用により、高度なサイバー攻撃活動が“スケール”（規模を大幅に拡大する。スケーラビリティを有する）することの問題やその対策の方向性について考察[4]する。

AI サービス各社による相次ぐ悪用事例報告

2026 年 2 月に Google が同社の Gemini を悪用した様々なサイバー攻撃活動に関する報告書を[5]公開した。Google 社（と買収された旧 Mandiant 社）は APT キャンペーンへの対応とアクターの追跡の知見が豊富であり、それぞれの悪用事象と APT アクターをはじめとした各脅威アクターとの紐づけに成功している。報告書で紹介された、各脅威アクターによる Gemini の悪用事例は以下の通りである。

標的の選定・偵察フェーズでの悪用：UNC6418、Temp.HEX（中国関連）

ソーシャルエンジニアリングの強化：APT42（イラン関連）、UNC2970（北朝鮮関連）

ツール開発ほか攻撃活動の支援：APT31（中国関連）、UNC795（中国関連）、APT42（再掲）

すでに APT アクターによる AI 活用が広く展開されている実態が明らかになっているが、後述の通り、APT のような高度な攻撃活動全体から見ると、AI 活用はあくまで補助的（イネーブラー）であり、一部の機能／リソースにおいて AI を活用しているように見受けられる。Google 社は、特にマルウェア開発において「革命的なパラダイムシフト」を起こすようなものではないが、「概念実証（PoC）型のマルウェアファミリーは、脅威アクターが将来のオペレーションにおいて AI 技術をどのように実装し得るかを示す初期の兆候」と指摘[6]している。

一方で、こうした限定的な AI 技術利用ではなく、AI エージェントを用いた、人的リソースの代替としての AI 利用が観測され始めている。2025 年 11 月に Anthropic 社が自社の AI サービス「Claude」を悪用した攻撃活動についてレポートを公開[7]した。この悪用は AI のエージェント機能を活用し、攻撃者の補助としてだけではなく、エージェント自体を攻撃の実行者として悪用したものであり、同社はサイバーセキュリティにおける「変曲点（inflection point）」であると評価している。具体的なアクター名は明示していないが、Anthropic 社は「高い自信を持って中国の国家支援グループと判定」しているとし、2025 年 9 月から大手テック企業、金融、化学、政府機関など 30 の組織に関する活動が観測されたと報告している。以下が報告された AI サービス悪用の各攻撃フェーズである。

フェーズ１：標的の選定・偵察

標的の選定はオペレータが手動で指示し、その後、Claude が複数の標的に対して同時並行で自律的な偵察を開始した。

フェーズ２：アタックサーフェス（侵入経路）選定

標的組織の IT インフラをマッピングし、認証メカニズムを分析、潜在的な脆弱性を特定した。

フェーズ３：脆弱性探索と検証

発見した脆弱性に応じたペイロード（悪意あるコード）を生成し攻撃テストを行い、標的システムからの応答を分析して当該脆弱性の悪用可能性を検証した。

フェーズ４：侵害拡大　横展開

オペレータからの承認を受け、Claude は標的ネットワーク全体の認証情報収集を行った。自動化された処理により、どの認証情報がどのサービスにアクセス可能か、各権限レベルとアクセス可能範囲のマッピングを行った。

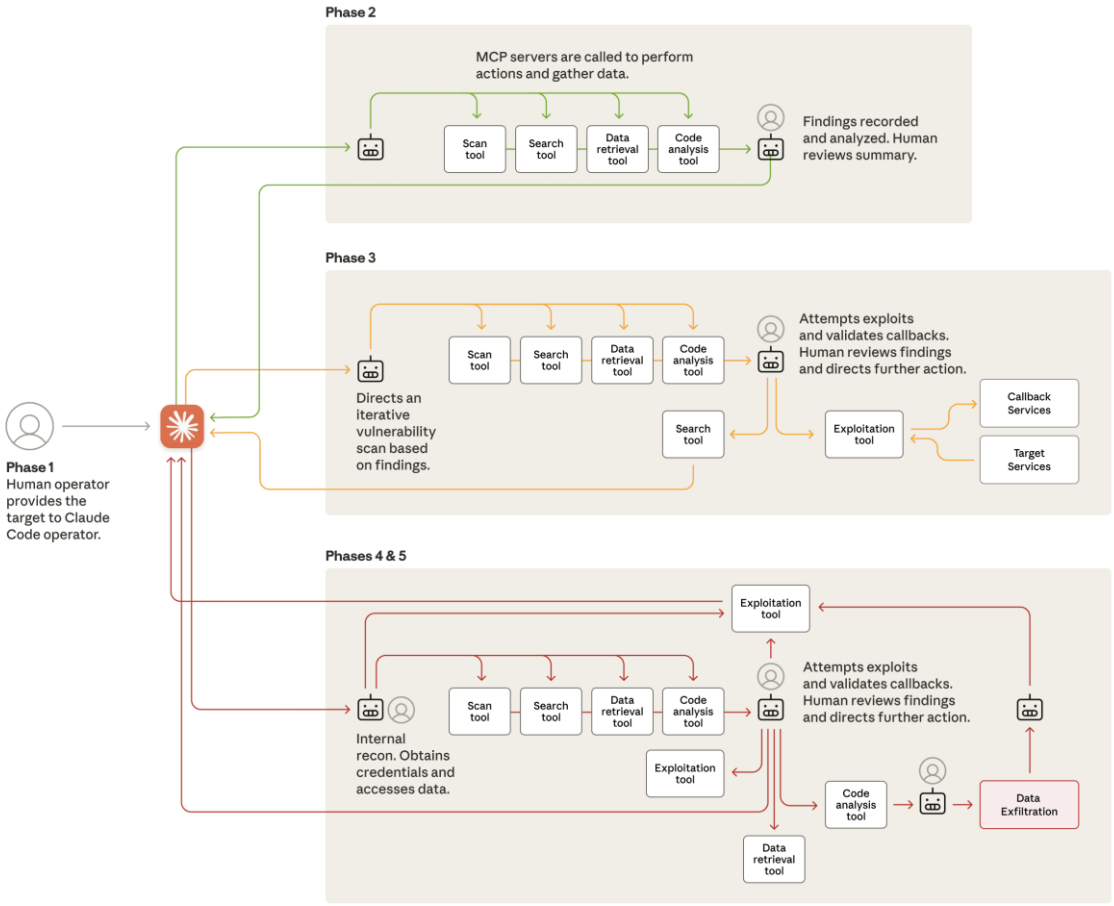
フェーズ５：窃取

標的組織のデータベース等へのクエリ実行、データ抽出、窃取したデータの整理などを行わせた。

フェーズ６：文書化

上記の全フェーズについて文書作成を行い、窃取した認証情報、データ、攻撃の過程などを記録した。この文書化により、各攻撃フェーズを分担するオペレータ間のシームレスな引継ぎが可能となる。

図：各攻撃フェーズについて Anthropic 社レポートからの抜粋（脚注 5 参照）



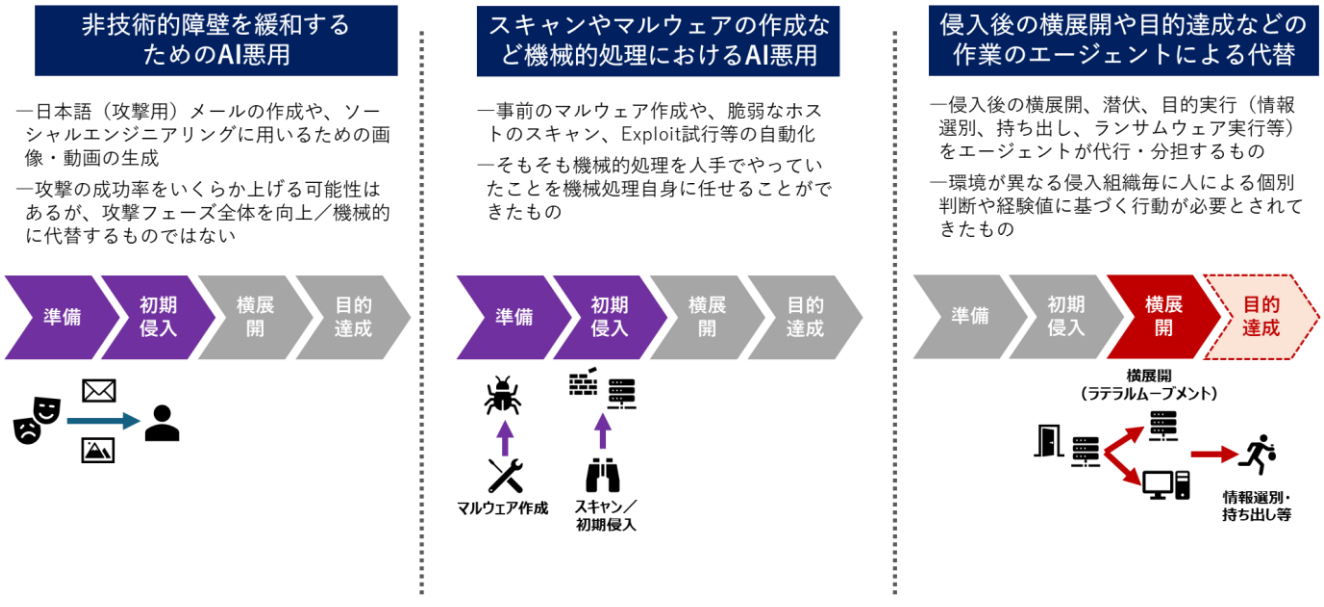
同社の 2025 年 8 月のレポートにおいて、いわゆる「バイブハッキング」事案が報告されているが、この事案では基本的に人間が攻撃の指揮を執る形で実施されていた[8]。今回報告の事案では自動化されたワークフローやオペレータの監督下で半自律的に攻撃が実行されている点で大きく異なっていると同社は指摘する。特にこれらの広範囲な攻撃キャンペーンを迅速・大規模に展開した点を、同社は「印象的だった」と評価している。他方で、筆者としては、特にフェーズ 4 の横展開（侵害拡大）やフェーズ 6（作業の記録・引継ぎ可能な状態）が自動化されている点に注目した。詳細は後述するが、高度なサイバー攻撃活動においては、侵入後のプロセスに高度な経験を有する人材が必要であるところ、こうしたフェーズが自動化・分業化できることで、従前人的リソースによって活動規模が制約されてきた高度なサイバー攻撃活動をスケール化することが可能になるのである。

アクター側の何が変わるのか

サイバー攻撃における AI の悪用により、「フィッシングメール／フィッシングサイトの日本語の巧妙化」、「ネットワークに面した機器への攻撃の自動化・範囲拡大」といった指摘が散見されるが、これらはもともと、AI 登場以前から従前のシステムである程度自動化がされてきたものであり、すでにスケールしたサイバー攻撃である[9]。また、2026 年 4 月に大きな話題となった、Claude Mythos による数千件の新たな脆弱性の発見（Project Glasswing）は APT のような高度な攻撃キャンペーンの初期フェーズの効率化・スケールを示している[10]。

他方で、前半に紹介した AI 悪用事例のインパクトが異なるのは、従前、大規模化が難しいと思われていた高度な APT キャンペーンが大規模化する点にある。本稿の問題意識は、AI が既に自動化されている初期侵入フェーズを大規模化・効率化する点よりも、従来は人的な熟練度に依存していた侵入後の横展開等の攻撃フェーズを大規模化しうるという点にある。

図：APT アクターによる AI 悪用のステップアップ（筆者作成）



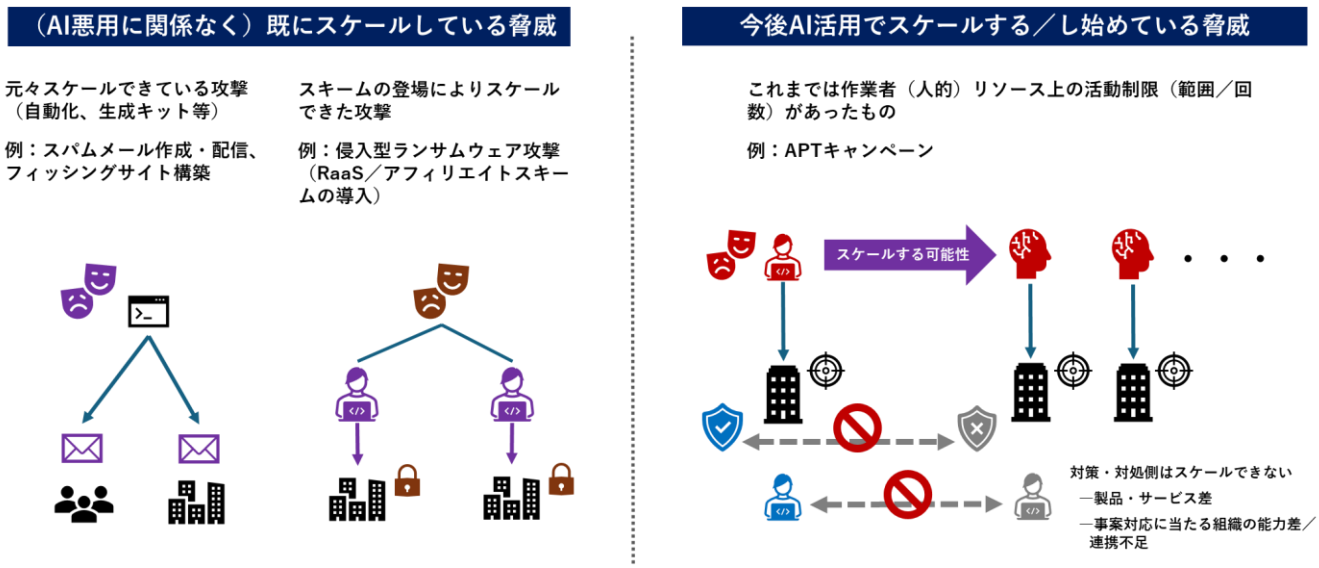
冒頭に紹介したメキシコの件は、国家を背景とするアクターによる、いわゆる APT キャンペーンではない可能性もあるが、Anthropic 社や Google 社が観測した事例は APT キャンペーンである。APT キャンペーンを行うためには高度なスキルと経験を持つ人材の確保が必要であり、特に中国関連では専門的知

見を持つ民間ハッカーやセキュリティ専門企業を利用した APT キャンペーンが相次いで明らかになっている。攻撃者（オペレータ）には、マルウェア開発やツール操作、脆弱性悪用などの技術的スキルだけでなく、特に侵入「後」の動きにおいては、標的組織内ネットワークの探索・全体像の把握、横展開のルート開拓、検出回避など、実践経験によって得られる知識も必要である。

初期侵入は基本的に同じ構成の対象機器を狙うため、自動化しやすい一方で、侵入後の標的組織内のネットワークというものは組織毎にその構成や運用状態、セキュリティ対策が大きく異なるため、「侵入して初めてその構造を探索して攻略する」必要があり、これまでは簡単に自動化できないと思われていた。だが、Anthropic 報告の攻撃活動はこの「侵入後の探索・侵害拡大」フェーズまでも自動化されているのである。この侵入「後」の自動化という新たな戦術は、これまでの APT アクターの活動範囲や頻度を縛っていた「(高度な）人的リソースの確保」の必要性を低減させ、今後大規模化する恐れがあると筆者は懸念するのである。

また、これも中国関連の APT 活動についてであるが、ここ 2〜3 年において、初期侵入、横展開、攻撃インフラ調達･運用をそれぞれ担うアクターが分業･連携する、マルチアクター戦術[11]が多用されており、活動全体の規模や攻撃頻度を高めることに成功しつつあると評価できる。AI はこの点でも活用（悪用）でき、前述の「フェーズ 6（作業の記録・引継ぎ可能な状態)」のように、他のユニット／担当者と分業・連携（引継ぎ）を容易にさせると考えられる。

図：AI 悪用による APT キャンペーンのスケール（筆者作成）



課題と対策

本稿で紹介した Anthropic、Google の両社とも、すでに生成 AI サービス提供における対策を講じている旨を表明している。事業者側での自主的な悪用防止措置はもちろん必要なものではあるが、攻撃活動への対処としては十分ではない。悪用防止策に対するあらたな回避手段が開発される可能性のほか、自主対策が不十分ではあるものの、高性能を有する新興 AI サービスの登場、さらには攻撃者やその支援組織・国家側が独自に AI サービスを開発[12]し、攻撃活動での利用を容認するようなケースが想定されるからである。

AI の悪用によって APT キャンペーンがスケールする恐れは、RaaS（ラース）スキームの拡散によってランサムウェア攻撃がスケールしたことと同様に考えられるだろう。2019 年以降、従前のランサムウェア攻撃に比べてより深く標的組織内部に侵入する、「侵入型ランサムウェア攻撃」が多発するようになり、世界的に爆発的な急増を見せたが、この背景には RaaS[13]と呼ばれる分業スキームがあり、RaaS によってスキルレベルが異なるアクターをランサムウェア犯罪に投入できようになったことで、ランサムウェア犯罪がスケールしたのである。

この時防御・対処側では、急増するランサムウェア攻撃に対して適切なインシデント対応・分析ができる専門組織（セキュリティ企業や専門機関等）が不足した。従前のランサムウェア攻撃と異なり、標的組織のネットワーク奥深くまで侵入し APT キャンペーン並みの戦術・ツールを使うケースもあり、インシデント対応における調査範囲・内容としては高度な APT キャンペーン被害事案の調査に近いレベルが求められることになった。他方でこうした調査経験がある専門組織はセキュリティ業界の中ではある程度限定されるため、急増する相談件数に対して十分な知見を持つ専門組織が不足する事態に陥った[14]のである。

AI 悪用による APT キャンペーンの大規模化によってこれと同じことがセキュリティ業界側で起きると筆者は懸念している。APT アクターによる AI 悪用がもたらす防御側のボトルネックは、「インシデント報告件数の増加」、「初動時間の猶予がさらに小さくなること」、「組織間調整の遅延・“摩擦”」、「専門的知見の需給バランスの不安定化」の 4 つである。こうした課題に対して、どのような対策が考えうるだろうか[15]。

〇サイバー攻撃「被害」の再定義と報告制度の見直し

AI 悪用により高度なサイバー攻撃キャンペーンがスケールすることで、行政機関への報告件数も増加す

ることが想定される。各国において、重要インフラ事業者を中心にインシデント報告の義務化や短期間での報告（48 時間～72 時間以内の初報など）を課す傾向[16]が強まっているところ、日本においては、令和 4 年 4 月からの改正個人情報保護法の施行により、個人データの漏洩について報告義務化がなされている。ランサムウェア攻撃の増加の影響もあるが、そのほか、報告義務化の周知が中小企業等まで広まったことで制度上の報告件数の増加[17]が続いている。そして、今年 10 月からは、「能動的サイバー防御」（ACD）整備に向けたサイバー対処能力強化法によって、基幹インフラ事業者に対してインシデント報告の義務化[18]が始まる。この制度においては、攻撃の初期フェーズで侵害されるようなネットワーク機器への侵害事象について、その認知時点での報告[19]を求めることから、従前報告対象となっていなかったインシデントの多くが報告対象として顕在化すると思われる。

前項で触れた通り、生成 AI 悪用によるサイバー攻撃の効率化については、目的達成までの全フェーズを効率化・高度化するにはまだいくつかのハードルがあるものの、その初期段階（ネットワーク機器の脆弱性発見・初期侵入等）を効率化させることは間違いない。前項で触れた「マルチアクター戦術」により、初期侵入経路の開拓に特化した分業アクターによる大規模なネットワーク機器侵害の攻撃キャンペーンが多発しているところ、AI 悪用によりさらに多くのアクターが同様の攻撃キャンペーンを展開できるようになった場合、サイバー対処能力強化法に基づく報告件数が急増することが予想される。行政機関へのインシデント報告の効率化として、報告様式の統一化[20]などが取り組まれ始めているところ、被害組織側の負担軽減がなされたとして、果たしてその膨大な報告情報を行政機関側が処理しきれるのかという新たな問題が登場する。AI 悪用により、高度なサイバー攻撃の「波状攻撃」「飽和攻撃」的な戦術[21]や、攪乱目的の大規模攻撃も懸念されるところ、大量の報告件数から最も警戒・対処が必要な事象を見つけ出すことができるのか、現状の制度ではその限界があるのではないかと筆者は考える。高度な APT キャンペーンのようなサイバー攻撃に限らず、インシデント報告数が急増することが想定される「AI 悪用時代」に即した、サイバー攻撃被害の再定義や報告制度の改善が必要である。

〇効率的なインシデント対応・調査手法の整備

攻撃側と同じく、防御側の様々なセキュリティ製品やサービスも AI 活用を進めており、自動化される多くの攻撃は十分に検出・防御可能になるであろう。他方で侵入を許してしまったあとのインシデント対応においては、被害組織毎にシステム構成・運用状況は異なり、また、インシデント対応では非技術的な対応のコスト（被害公表・広報対応、顧客・取引先などの対応、行政機関への報告等）が発生するため、（少なくとも現時点では）人間がその多くに対応せざるを得ない。速やかに侵入を検知、あるいは攻撃被害に遭う前に注意喚起等の情報や対策手段を得ることができたとしても、組織内外の対応に手間取るこ

とで被害予防・被害拡大防止の機会を失ってしまうのである。

サイバー攻撃被害として公になるものは、情報漏洩やサービス停止など実際の被害が確認された場合であり、侵入を防げた場合、または侵入されたものの、すぐに検出し追い出すことができた場合などは通常明らかにされていない。「侵害されたものの、早期に検出・対応できた」案件も相当数存在している。この初期対応フェーズで課題となるのは、攻撃者側の AI 導入により、「侵害されてから“迎撃”できるまでの時間（機会）」が極めて短くなることが予想されるため、これまで以上に検知から対応へのスピードアップが求められることになる。組織内の調整、行政機関等とのやりとりのコストといった、インシデント対応における技術的対応以外のコストについて、可能な限り無駄を省き、被害組織が初動段階で技術的対応に集中できる環境整備[22]が必要である。

〇情報開示・情報共有

APT キャンペーンが大規模化した場合、攻撃を広範囲に行うほど、検知される確率も高くなる。他方で、被害現場間で情報が共有されなければ、その他の被害は放置されることになる。

AI を利用しているかどうかは不明であるが、2024 年と 2025 年に複数回にわたる大規模な攻撃キャンペーンが発覚した UNC5221 による Ivanti 製品の脆弱性を狙った攻撃キャンペーン[23]では、数千台以上の同社製機器を狙ったグローバルな攻撃が展開されたため、いずれの攻撃の波も 1 か月以内には捕捉・対処がなされている。この際には最初のゼロデイ攻撃を発見した Google／Mandiant 社が速やかに分析レポートを開示[24]するとともに、脆弱性が悪用された Ivanti 社をはじめ、各国の専門機関等が注意喚起を実施した。

広範囲な攻撃であるほど、攻撃キャンペーンを捕捉できる可能性は高くなり、また、各被害現場での調査が早く進むほど、現場の証拠（アーティファクト）から得られる情報量も増え、攻撃の全容解明や被害拡大防止につながりやすくなる。問題はこうした動きが人・組織間の摩擦で阻害されることである。最初に被害に気づけた組織からの情報が共有されなかったり、原因となった製品メーカーからの情報開示が遅れる[25]など、非技術的な要因によって社会全体の対応スピードが遅延することで、攻撃キャンペーンを捕捉できないまま、事後対応に陥りがちである。

〇対処能力の“見える化”

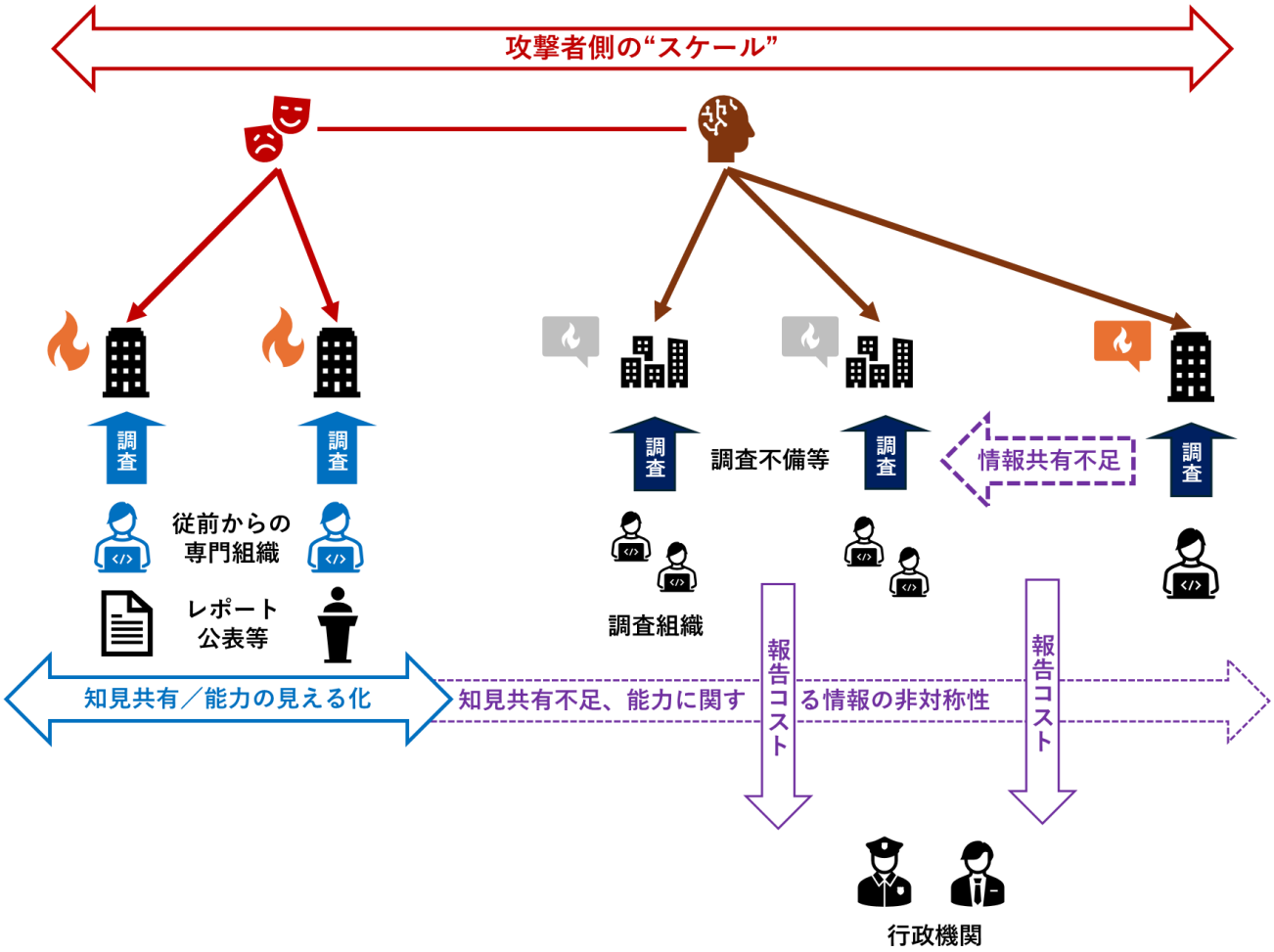
先述の通り、ランサムウェア攻撃の増加フェーズにおいては、技術・知見的に適切に対処可能な専門組

織が不足する事態に陥った。さらに、専門組織が少ない地方において「頼る先がない」被害組織がとにかく相談できそうな先に依頼することでミスマッチが起きていた。

APT キャンペーンがスケール化した際も同様の現象が起きるのではないかと考えられるところ、セキュリティ業界側全体の知見向上もさることながら、利用者（被害組織）が正しく専門組織を選択できる仕組みも必要である。個別のインシデント対応は NDA（秘密保持契約）などにより情報が制約されており、また、被害公表で開示される情報も限定的であるため、「どの専門組織がどのような作業ができ、どの程度の専門知見を有しているのか」外部評価を受けにくいという構造[26]にある。

他方で、特に APT キャンペーンにおいては専門組織から分析レポートが公表されることが多く、こうした開示情報を通じて当該専門組織（や担当したアナリスト）の能力を評価することが可能[27]となっている。こうした情報開示により利用組織（被害組織）が知見のある専門組織を選びやすくなるだけでなく、業界内においても相互評価のメカニズムが働き、また、知見の共有が広く行われるようになるため、業界全体でのスキルの底上げやミスマッチの回避が進むと思われる。

図：アクター側のスケールに対して我が方側の課題／非対称性（筆者作成）



さいごに

APT キャンペーンのような高度なサイバー攻撃における AI の悪用がもたらす影響について未だ断定的なことは言えないだろう。Google が報告した事案では AI 活用はまだ限定的なものであり、試行錯誤がなされていることが垣間見られる。また、Anthropic が報告した事案でも AI 側のハルシネーションが起きている点が指摘されており、攻撃の有効性にはまだ疑問が残る状況である。前述の通り、攻撃の範囲が広く活発に行われれば、いずれかの標的組織やセキュリティ製品が早期検出できる機会が増える点も指摘できる。さらに、「人・組織同士の摩擦を減らす」ことで、大規模化しようとする APT 活動にも対処できる防御側の改善の余地がまだ残されている。

2026 年 4 月に英 AISI（AI Security Institute）が Mythos Preview のサイバー攻撃能力評価に関するレポートを公表した[28]。これは、単独の問題を解くようないわゆる CTF（Capture The Flag）方式のテストだけ

ではなく、実際の標的組織のネットワーク環境模した「サイバーレンジ」上での多段階の攻撃シナリオのシミュレートが含まれている。Mythos Preview はやはり従前の AI モデルよりも好成績を収め、攻撃フェーズのほとんどを実行することに成功しているが、このテストでは実際のネットワーク環境で実装されているような能動的な検知手段などは用意されていないことから、AISI 側も「よく防御されたシステムを攻撃できるかは確証を持って断言できない」とし、そうした実運用環境により近い環境での検証が今後の課題であるとしている。

AI 悪用によるサイバー脅威の変化については、抽象的な話題が多く、具体的事例や個別類型／アクター別の考察がほとんどなされていないのが現状である。特に APT キャンペーンについては、サイバー攻撃／犯罪全体の規模からすればかなり限定された範囲の事象であり、また、被害現場対応を実際に経験できるのはインシデント対応をサービスとして有し、専門的知見を持った一部の専門組織に限定されるため、「語れる」関係者が比較的少ない分野である。本稿では筆者が実際に本務側で対応にあたってきたランサムウェア攻撃の急増と、これまで分析・追跡してきたマルチアクターモデルの観点から、APT キャンペーンにおける AI の悪用によって防御側が直面する可能性のある課題を整理した。

APT 事案対応のような高度なインシデント対応は技術的な対応の難しさだけでなく、被害の社会的影響から、被害公表・メディア対応、行政機関への報告・届出などの非技術的な対応コストが極めて高い事案となる。こうした APT キャンペーン対応に日々あたっている脅威アナリストたちはインシデント対応における様々な人的・組織的・制度的な“摩擦”を目の当たりにしており、なおさらアクター側の“スケール”の脅威への懸念を募らせているのである。本稿を通じてそうした危機感の一部が伝われば幸いである。

[1] Bloomberg, “Hacker Used Anthropic’s Claude to Steal Sensitive Mexican Data”, https://www.bloomberg.com/news/articles/2026-02-25/hacker-used-anthropic-s-claude-to-steal-sensitive-mexican-data. なお、この事案で使われたとされる Claude Code は 2026 年 4 月に大きな話題となった Cloude Mythos ではもちろんない。詳細等は脚注 2 を参照。

[2] Gambit Security, “A Single Operator, Two AI Platforms, Nine Government Agencies: The Full Technical Report”, April 10,2026, https://gambit.security/blog-post/a-single-operator-two-ai-platforms-nine-government-agencies-the-full-technical-report.

[3] セキュリティベンダ等の四半期レポートなどにおいて、攻撃／戦術の傾向として AI 悪用が取り上げられるようになったほか、個別の APT キャンペーンに関する分析レポートにおいても AI 悪用事例が登場するようになってきている。例：北朝鮮関連の APT グループ Konni が用いていた Powershell バックドアに AI を使った作成痕跡が見つかった例（https://research.checkpoint.com/2026/konni-targets-

developers-with-ai-malware/）

[4] 本稿は 2026 年 2 月に防衛研究所にて行った NTT セキュリティ・ジャパン株式会社のアナリストとの研究会で討論した内容を元に企画したものである。特に AI 悪用による「スケール」の視点は同社の羽田大樹セキュリティプリンシパルから示唆をいただいたものである。

[5] Google, “GTIG AI Threat Tracker: Distillation, Experimentation, and (Continued) Integration of AI for Adversarial Use”, February 13, 2026, https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use?hl=en.　なお、2026 年 5 月 12 日に Google が新たな AI 悪用に関するレポートを公開しているが、本稿掲載準備の都合から当該レポートの精査・言及まで至れていない点はご容赦いただきたい。またの機会に触れることとしたい。Google, “GTIG AI Threat Tracker: Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access”, May12,2026, https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access?hl=en

[6] 前掲注 5

[7] Anthropic, “Disrupting the first reported AI-orchestrated cyber espionage campaign”, November 13,2025, https://www.anthropic.com/news/disrupting-AI-espionage.

[8] Anthropic, “Threat Intelligence Report:

August 2025”, https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf.

[9] AI 悪用によってサイバー脅威がどのように変わるかについては、まだその議論が本格化していない。AI 利用によって広範囲・高速化する可能性については、防御側製品・サービスの AI 利用による高度化とどう拮抗するのか、あるいは力負けする可能性があるのか議論がまだ少ない。また、サイバー攻撃者による AI 悪用によって「日本語の壁」が突破される、といった言及が散見されるが、無差別なフィッシングメールやフィッシングサイトはともかくとして、ネットワーク内部まで侵入するような攻撃の場合、ネットワーク環境、システム環境を侵害する作業に日本語は基本的に必要ない。本稿では APT キャンペーンに考察対象を絞ったが、具体的な攻撃類型／アクター毎での具体的な考察が求められている。

[10] Anthropic, “Project Glasswing”, https://www.anthropic.com/glasswing, “Assessing Claude Mythos Preview’s cybersecurity capabilities”, https://red.anthropic.com/2026/mythos-preview/.

[11] HarfangLab, “Further insights into Ivanti CSA 4.6 vulnerabilities exploitation,” https://harfanglab.io/insidethelab/insights-ivanti-csa-exploitation/; Cisco Talos, “Redefining IABs: Impacts of compartmentalization on threat tracking and modeling,”https://blog.talosintelligence.com/redefining-initial-access-brokers/.

[12] 先述の Google の報告書では、Gemini の学習モデルを“盗もう”とする、知識蒸留（knowledge distillation）の試みを阻止している点も報告されている。また、当初から悪用前提でアンダーグラウンドマーケット等に流通しているものも登場している。Unit42, “The Dual-Use Dilemma of AI: Malicious LLMs”, November 25, 2025, https://unit42.paloaltonetworks.com/dilemma-of-ai-malicious-llms/.

[13] RaaS（Ransomware as a Service）：スキームのオーナー（オペレータ）がランサムウェアを実行者（アフィリエイト）に貸し出し、収益（身代金）の一部のその利用料としてせしめる犯罪スキーム。身代金交渉用のインフラ・交渉窓口の提供も有しており、また、アフィリエイトに対して攻撃手順等を記したマニュアルを配布するなどしており、自前でランサムウェアを用意できないアクターや低スキルのアクターでもインパクトの大きい侵入型ランサムウェア攻撃を実施可能なため、アンダーグラウンドマーケットを通じて大規模に人的リソースが流入し、一気に“スケール”してしまったのである。

[14] JPCERT/CC 佐々木勇人「ランサムウェア攻撃のアクター特定をすべきこれだけの理由」、JSAC2024、https://jsac.jpcert.or.jp/archive/2024/pdf/JSAC2024_2_6_hayato_sasaki_jp.pdf。

[15] 2026 年 4 月の Anthropic 社の Claude Mythos による Project Glasswing のケースのように、防御側が生成 AI を活用することで未発見の脆弱性を修正することが可能である。このほか、様々なセキュリティ製品・サービスでの AI 活用が始まっているが、防御側における AI 活用は本稿の主たるトピックではないため特に触れていない。

[16] PwC「サイバー攻撃被害に係る公表」に関する国内組織実態調査 第 2 回」、2025 年 3 月 10 日、https://www.pwc.com/jp/ja/knowledge/thoughtleadership/cyber-attack-survey2024.html。

[17] 個人情報保護委員会「令和 6 年度個人情報保護委員会年次報告書」、https://www.ppc.go.jp/files/pdf/070610_annual_report.pdf。

[18] 法律の概要及び基幹インフラ事業者によるインシデント報告の義務化に関する解説資料。内閣官房国家サイバー統括室、「サイバー対処能力強化法及び同整備法について」、https://www.cas.go.jp/jp/seisaku/cyber_anzen_hosyo_torikumi/pdf/setsumei.pdf。

[19] 国家サイバー統括室「「重要電子計算機に対する不正な行為による被害の防止に関する法律に基づく特別社会基盤事業者による特定侵害事象等の報告等に関する命令案」に関する意見の募集について」、https://public-comment.e-gov.go.jp/servlet/Public?CLASSNAME=PCMMSTDETAIL&id=095260180&Mode=0。

[20] 行政機関への対応コストにおいて問題視されているうちの１つが「複数機関へのそれぞれ異なる様式での報告」の対応コストである。現時点ではランサムウェアと DDoS 攻撃の 2 類型に限定されているが、行政機関への報様式については、その一本化が図られることとなった。

国家サイバー統括室「サイバー攻撃による被害発生時のインシデント報告様式の統一について」、2025 年 10 月 1 日、https://www.cyber.go.jp/policy/group/cyber/yoshikiichigenka.html。

[21] 2026 年 4 月に行われた米陸軍と民間企業による「AI サイバー戦」演習（AI TTX 2.0）を踏まえ、陸軍主席サイバー顧問の Brandon Pugh 氏は記者らに対して「敵が AI を駆使して単発の決定的なサイバー攻撃を仕掛けるのではなく、米陸軍の防衛体制に絶えず適応しながら波状攻撃を繰り返す」シナリオが想定されたと言及した。U.S. Army Public Affairs, “Army convenes industry leaders for AI tabletop exercise focused on cyber defense”, US Army, May 4,2026, https://www.army.mil/article/292158/army_convenes_industry_leaders_for_ai_tabletop_exercise_focused_on_cyber_defense; Chris Panella「米陸軍が「AI サイバー戦」演習を実施。敵の AI は凄まじい速さで、システムの脆弱性を突いてきた」、BUSINESS INSIDER、2026 年 5 月 4 日、https://www.businessinsider.jp/article/2605-how-us-army-is-readying-for-enemy-ai-cyberspace-attack/。

[22] 前掲注 20

[23] 2024 年末から 2025 年にかけての攻撃キャンペーンに関する JPCERT/CC からの注意喚起や分析レポートは以下のとおり。2025 年 1 月～、一般社団法人 JPCERT コーディネーションセンター（JPCERT/CC）早期警戒グループ「Ivanti Connect Secure などにおける脆弱性（CVE-2025-0282）に関する注意喚起」、JPCERT/CC、2025 年 2 月 13 日、https://www.jpcert.or.jp/at/2025/at250001.html；増渕維摩「Ivanti Connect Secure の脆弱性を利用して設置されたマルウェア SPAWNCHIMERA」、JPCERT/CC Eyes、https://blogs.jpcert.or.jp/ja/2025/02/spawnchimera.html。

2025 年 4 月～、一般社団法人 JPCERT コーディネーションセンター（JPCERT/CC）早期警戒グループ「Ivanti Connect Secure などにおける脆弱性（CVE-2025-22457）に関する注意喚起」JPCERT/CC、2025 年 4 月 30 日、https://www.jpcert.or.jp/at/2025/at250008.html；増渕維摩「Ivanti Connect

Secure の脆弱性を起点とした侵害で確認されたマルウェア」JPCERT/CC Eyes、2025 年 7 月 18 日、https://blogs.jpcert.or.jp/ja/2025/07/ivanti_cs.html。

[24] Google/Mandiant による 2024 年末から 2025 年にかけての攻撃キャンペーンの分析レポート

2025 年 1 月　John Wolfram, Josh Murchie, Matt Lin, Daniel Ainsworth, Robert Wallace, Dimiter Andonov, Dhanesh Kizhakkinan, Jacob Thompson, “Ivanti Connect Secure VPN Targeted in New Zero-Day Exploitation,” Mandiant, January 9, 2025, https://cloud.google.com/blog/topics/threat-intelligence/ivanti-connect-secure-vpn-zero-day/?hl=en.

2025 年 4 月　Google/Mandiant, “Suspected China-Nexus Threat Actor Actively Exploiting Critical Ivanti Connect Secure Vulnerability (CVE-2025-22457)”, April 4, 2025, https://cloud.google.com/blog/topics/threat-intelligence/china-nexus-exploiting-critical-ivanti-vulnerability?hl=en.

[25] 例えば、脆弱性情報の開示をめぐる人・組織間の“摩擦”については、佐々木勇人「解説：脆弱性関連情報取扱制度の運用と今後の課題について（前編）～公益性のある脆弱性情報開示とは何か～」JPCERT/CC Eeys、https://blogs.jpcert.or.jp/ja/2025/09/handling_vul_info_1.html。

[26]「経済産業省産業サイバーセキュリティ研究会 サイバー攻撃による被害に関する情報共有の促進に向けた検討会 報告書」2024 年 3 月、https://www.meti.go.jp/shingikai/mono_info_service/sangyo_cyber/cyber_attack/pdf/20231122_2.pdf。

[27] こうしたレポート公表は専門組織の PR でもあり、また、担当したアナリストの成果発表でもあるため、積極的に公開されている。こうした情報開示は情報共有の枠を超え、グローバルな情報共有にもなっており、特定の APT アクターの攻撃活動を世界各国の専門組織／アナリストが追跡するための情報の基盤となっている。最近はその数は減ったものの、侵入型ランサムウェア攻撃が急増した当初は事案対応にあたって専門組織が多くの分析レポートを公表し、専門組織／アナリスト側の知見共有に貢献してきたのである。

[28] “Our evaluation of Claude Mythos Preview’s cyber capabilities”, AISI, April 13, 2026, https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities. ブログ記事内に論文本体へのリンクが掲載。

PROFILE

佐々木　勇人

先進領域研究部サイバー安全保障研究室特任研究員（本務先：一般社団法人 JPCERT コーディネーションセンター　脅威情報アナリスト　サイバーセキュリティコーディネーショングループ担当部門長　兼　政策担当部長）

専門分野：能動的サイバー防御（ACD）、サイバー攻撃者のアトリビューション、サイバー脅威インテリジェンス、サイバー情報共有と攻撃被害公表

防衛研究所企画部企画調整課
直　通：03-3260-3011
代　表：03-3268-3111（内線 29177）
防衛研究所 Web サイト：www.nids.mod.go.jp