



フロンティア AI の登場による

企業のサイバーセキュリティリスクの変化と対応

山根 晴貴（ビジネスリスク本部 研究員）

要約

AI 技術の急速な発展が、サイバーセキュリティを取り巻く環境に大きな変化をもたらしている。特に「Claude Mythos」シリーズに代表されるフロンティア AI 技術は、人間をはるかに上回る速度と精度で脆弱性の発見と悪用を行うことを可能にした。AI との直接的な因果関係は明らかではないものの、2026 年夏以降、大規模な不正アクセス被害が続発していることから、環境変化の重大性がうかがえる。他方で、AI の適切な利活用は、企業のセキュリティ強化に繋がる重要な鍵でもある。本稿では AI の影響による近年のサイバーセキュリティ領域の動向変化を整理した後、企業に求められる対策として①AI の利用全般に係る適切な人間の関与および統制の整備、②AI を活用したセキュリティ（AI for Security）への迅速な転換を可能にする体制・ルールおよびインテリジェンス態勢の構築、③AI を利用する際のセキュリティ（Security for AI）、特に AI エージェント製品の導入に向けた備えの三点を整理する。

目次

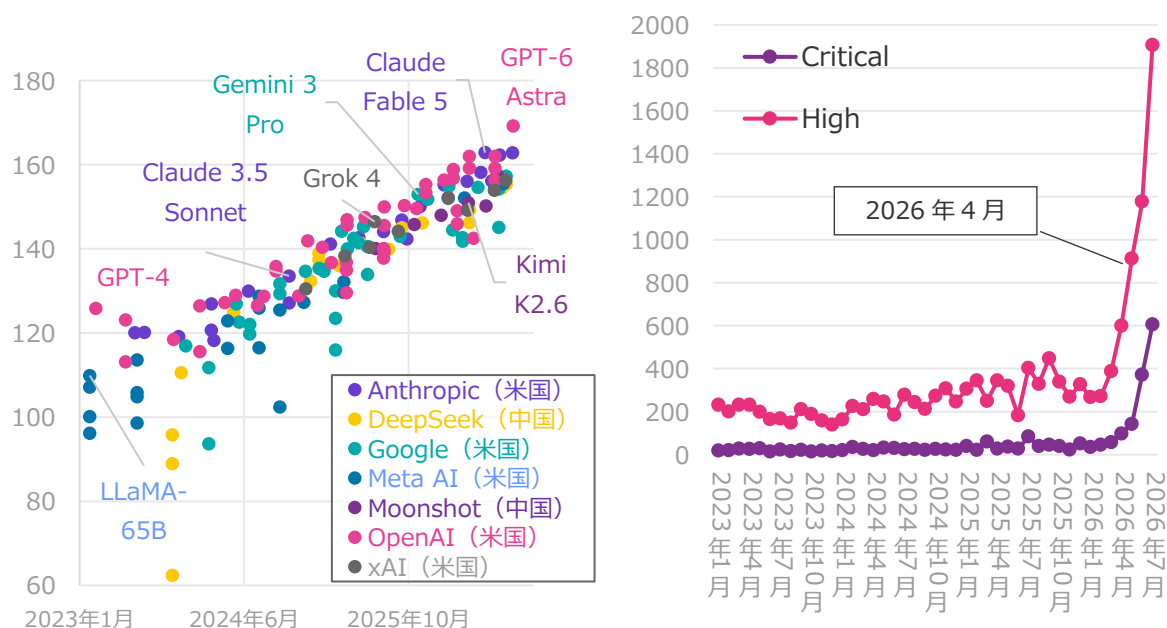
1. AI を巡る動向の概観	2
2. AI によるサイバーセキュリティの環境変化	3
(1) 攻撃側の変化	3
a. 攻撃の質の向上	3
b. 攻撃の量の増加と高速化	4
(2) 防御側の変化	5
a. 「AI for Security」の急速な拡大	6
b. 「Security for AI」領域の登場	6
3. 企業に求められる対応	6
(1) Human in Command：人間による AI の適切な制御	7
(2) AI for Security：自動化された防御への転換	8
a. 体制	8
b. ルール	8
c. インテリジェンス	8
(3) Security for AI：「自律性」リスクへの備え	10

1. AI を巡る動向の概観

2022 年 11 月の ChatGPT-3.5 公開以降、AI 技術は、主にテキストや画像、音声等を生成する「生成 AI」を中心に発展してきた。しかし、近年はサイバーセキュリティに直結する性能に特化した Claude Mythos シリーズ等の「フロンティア AI¹」や、自律的に判断・行動しながら一連の業務を遂行する「AI エージェント」など、新たな AI 技術の発展・浸透も目覚ましい。本稿の主たる焦点はフロンティア AI の登場が企業のサイバーセキュリティに及ぼす影響であるが、議論の性質上、従前の生成 AI 全般や、AI エージェントについても、一部の項で併せて取り上げることとする。

汎用的な AI の処理能力の向上は、各トップランナー企業の競争を通じ、比較的一定のペースで進んでいる。他方、サイバーセキュリティ関連のタスクに関する能力は急激な進化を遂げた可能性がある。公開データからは、主要な IT 製品に関する重大な脆弱性の発見件数が、“Claude Mythos Preview”の限定公開がなされた 2026 年 4 月以降に爆発的な上昇を示していることがわかる²（図 1）。Claude Mythos へのアクセスは現時点では審査制であり、利用可能な企業は一部に限定されているものの、今後同水準のサイバー攻撃能力を備えた AI が、一般アクセス可能な形で開発・公表される可能性は高い。したがって、フロンティア AI への直接的なアクセス権の有無にかかわらず、すべての企業は AI の登場によってサイバーセキュリティ領域に生じている変化に備える必要がある。

■ 図 1 AI ベンチマークの進化（左）と、重大な脆弱性の発見件数（右）



左図：主要企業の AI モデル公開年月（横軸）と Epoch Capabilities Index (ECI)スコア（縦軸）。“Epoch Capabilities Index (ECI),” [Epoch AI](#) (updated September 8, 2026) を基に筆者作成。ECI は非営利研究機関 Epoch AI が作成・公開している AI のパフォーマンス評価指標。複数の分野・難易度の AI ベンチマークを統合した指標であり、幅広いモデルのパフォーマンス能力を長期的なタイムスパンで共通比較することに長ける。

右図：各月（横軸）に報告された、Critical、High に分類される Common Vulnerabilities and Exposures (CVE) の件数（単位：件）。Luke Emberson, “Serious cyber vulnerability disclosures kept climbing in July,” [EPOCH AI](#) (July 31, 2026) を基に筆者作成。なお、件数は AWS, Apache, Apple, Cisco, Google, Linux, Microsoft, Mozilla, NVIDIA, Oracle, Red Hat, Adobe, IBM, Intel, AMD, Qualcomm, Samsung, SAP, VMware, GitHub, OpenSSL 社製品についてのものであり、脆弱性の報告手順やラベル付け基準、報告頻度は組織によって異なる。

¹ 「フロンティア AI」に公的な定義は存在しないが、本稿では 2026 年 10 月現在の最先端・最高水準の AI 技術、という意味でこの語を用いる。

² ただし、脆弱性発見件数の急増と“Claude Mythos Preview”との間の因果関係の有無は、公開情報からは不明である。

2. AI によるサイバーセキュリティの環境変化

AI は攻撃側・防御側の双方を利する技術革新である。現時点では、サイバーセキュリティ能力に特化したフロンティア AI のトップランナーは米国企業（特に Anthropic 社、OpenAI 社）であり、それらモデルへのアクセスは、攻撃者に悪用されないよう厳格に管理されている。そのため、利用できる技術水準の高さという観点では、フロンティア AI を用い脆弱性診断やパッチ作成を行うことができる IT ベンダやセキュリティ技術企業に優位性があると言えよう。ただし、中国製 AI 製品の性能急伸やオープンアクセスモデルの性能向上を踏まえると、上記の形勢は必ずしも今後も維持されるとは限らない。

他方、攻撃者と、フロンティア AI へのアクセス権を持たない一般ユーザ企業との間では、攻撃側に優位な構造変化が生じている可能性がある。自動化・効率化によるサイバー攻撃の“生産性”向上は、攻撃者に即座に利益をもたらす。他方、企業にとっては、新たなシステムの導入はコストである。さらに、攻撃者には既存の業務プロセスや組織内の承認手続きといった制約が少なく、AI 技術を素早く攻撃に取り入れることができるが、企業が新システムを導入するにあたっては、有効性の検証、コストパフォーマンスの分析、社内承認手続き、そして導入後の説明責任の確保といった、ビジネス運営上の様々な制約を課される³。これらの全体像に基づきながら、本章では、サイバー攻撃を行う攻撃者と、サイバー攻撃から自社の資産を守る企業の双方の観点から AI の登場による技術や戦略の変化を整理する。

(1) 攻撃側の変化

AI は、サイバー攻撃を行う主体のような悪意ある利用者にとっても有用なツールとなる。本項では、AI の登場により生じている代表的な脅威トレンドを紹介する。攻撃者は、AI を用い攻撃の「質」「量（速度）」の双方を劇的に高めていると考えられる。無論、「質」と「量」は完全に独立した観点ではなく、攻撃の「質」すなわち攻撃の成功率（検知・防御の難易度）や侵害の深さ・大きさと、「量」すなわち攻撃の遂行速度や総件数は相互に関連するものである。攻撃の遂行速度は成功率の重要なファクターであるし、攻撃の総量が増加しセキュリティ業者やフォレンジック業者への対応・調査依頼が逼迫すれば、被害企業が必要な措置を迅速に受けられず被害が甚大化しやすくなる。

a. 攻撃の質の向上

AI が攻撃の高度化に利用されているひとつの例は、マルウェアの難読化・検知困難化である。一般的なオープンアクセスモデルには「コンピューターウイルスを作って」といった悪意ある依頼は拒絶するよう安全弁が施されている。しかし、「研究用」等の虚偽の目的とともに伝える、ないしそれがマルウェアであることを伏せたうえで「このコードを、同じ挙動を保ったまま、複雑にして」等といった婉曲な依頼を送れば、実行してしまうケースがある。すでに、AI を用いてマルウェアのコードを煩雑化し悪意ある振る舞いをする箇所を隠したり、正規のソフトウェアの挙動を模倣させたりすることで、対策ソフトウェアによる攻撃検知を回避する手法が確認されている。より発展的な脅威は、エージェント型 AI 搭載の適応型マルウェアの登場である。侵害の実行中に自己を変化させ自律的に検知や隔離をかくぐるこのようなマルウェアは、従来よりもはるかに短い時間の攻撃で、甚大な被害をもたらす可能性がある⁴。

³ 本質的には、この非対称性は AI の導入のみならず、サイバーセキュリティ全体の構造である。攻撃側にとって技術や人材への投資は利益（プラス）を増加させるものであるのに対し、防御側にとってセキュリティへの投資は損失（マイナス）を小さくするために支払うべきコストである。また、攻撃側は多数の攻撃を仕掛け一度（一社）でも攻撃に成功すれば利益が生じるのに対し、防御側は一度でも自社が防御に失敗するだけで、莫大な損失を被る。

⁴ 悪意を隠蔽し、工夫した指示文を用いれば、ChatGPT のようなオープンアクセスモデルをマルウェア作成や難読化、配布等の目的で利用できることは、生成 AI が普及して比較的時間もない時期（2023 年初頭頃）にすでに専門家の実践等により明らかになっていた。吉川孝志『マルウェアの教科書 増補改訂版』（2023）、日経 BP、207-210 頁。

生成 AI によるソーシャルエンジニアリング攻撃の高度化も重要な脅威である。ソーシャルエンジニアリング攻撃とは、人的な要因（油断や信頼、恐怖心、無知等の心理的な隙）を利用した攻撃全般を指し、偽メール等を通じて情報を詐取するフィッシングが典型例である。近年は非日本語圏の攻撃者であっても、生成 AI を活用すれば巧妙なメール文を作成することができる。さらに、インターネット上から標的企業や個人に関する情報を収集し相手ごとにアレンジした文面を作成することも、AI を利用すれば低コストで実行可能である。セキュリティ専門家を対象にした調査では、AI によってもたらされる「最大の脅威」として「高度に個人化されたフィッシング」が最多の回答を集めた（表 1）。今後は、音声・映像の生成技術（ディープフェイク）を組み合わせ経営層や取引先を装い、電話・オンライン会議を通じて送金や情報漏洩を誘導するといった、より高度な詐欺も現実的な脅威となってくるであろう。

■ 表 1 セキュリティ専門家が挙げる AI によるサイバーセキュリティの「最大の脅威」

脅威	回答割合	脅威の概要
高度に個人化されたフィッシング	50%	AI 生成された自然な文章と、標的の属性や興味関心に沿った内容で成功率を高めたフィッシング攻撃。
自動化された脆弱性スキャンとエクスプロイト連鎖	45%	標的のもつ脆弱性を偵察し、それを悪用し侵入するための足掛かりや被害拡大の手段（エクスプロイト）を自動で作成する。
適応型マルウェア	40%	侵害中に自己を変化させ、自律的に検知や隔離を回避する。
ディープフェイクによる音声詐欺	40%	経営層や取引先の声紋を生成 AI で再現し、詐欺を行う。

※調査対象は各国・業界における企業のCISO、IT部門長、情報管理者、実務者。「最大の懸念にあたる項目（as one of their biggest worries）」に該当する項目を回答（複数選択可）。サンプル数は1,540名。

出典：“The State of AI Cybersecurity 2026 The Threat Landscape,” [DARKTRACE](#), (February 3, 2026) を基に筆者作成

b. 攻撃の量の増加と高速化

質的な変化と並行して、AI 技術は攻撃の「量」および「速度」の面においても、大きな変化をもたらしている。

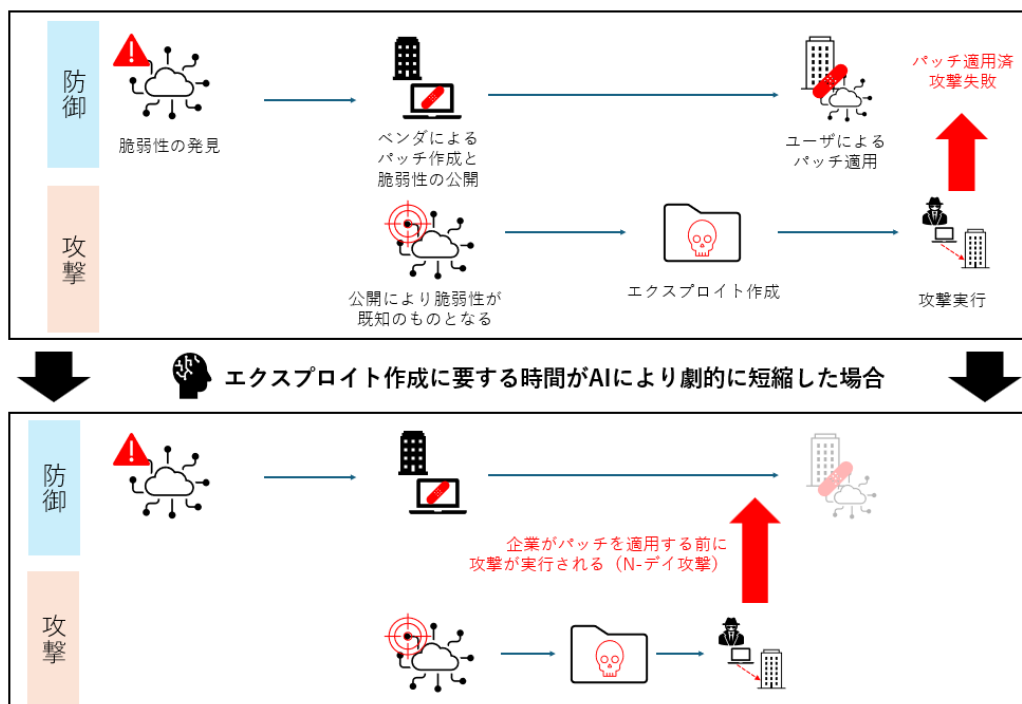
最も重大な変化は、脆弱性発見からエクスプロイト（脆弱性を悪用するための具体的な攻撃プログラム）の作成、攻撃実行に至るまでに要する時間の劇的な短縮であろう。従来、脆弱性の発見とエクスプロイトの作成には専門的な知識と相応の時間・労力を要するということは、攻撃者・防御者の共通認識であった。しかし、フロンティア AI は人間をはるかに上回る能力で脆弱性を発見し、エクスプロイトを作成することができる⁵。これは、サイバーセキュリティの根本を覆しうる技術革新である。例えば、毎月第 2 火曜日にパッチ（脆弱性に対処するための修正プログラム）をリリースする Microsoft 社の「パッチチューズデー」に代表される月次パッチ適用の慣習は、攻撃者が脆弱性に気付きエクスプロイトを作成するまでに基本的には 30 日以上を要するという前提に基づいたものであった。仮に誰もがフロンティア AI を用いることができれば、攻撃者はベンダ企業とほぼ同時に脆弱性に気付き、かつ僅か数時間の間にエクスプロイトを作成してしまう可能性がある。そのような状況ではもはや従来のパッチ適用のタイムスパンでは防御が全く間に合わず、ベンダによるパッチ開発前の攻撃（ゼロデイ攻撃）や企業のパッチ適用前（N-デイ攻撃；図 2）の発生可能性が劇的に高まる。Claude Mythos シリーズをはじめとするフロンティア AI へのアクセス権が現時点では一部企業に制限されている背景には、おおよそ上記のような理由がある。

適応型マルウェアについては、Google の脅威インテリジェンスチームにより、侵害の実行中に「ジャストインタイム」で AI を用いスクリプトの生成や難読化を行うマルウェアの一群の存在が報告されている。” GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools,” [Google Threat Intelligence Group](#) (November 6, 2025)。

⁵ Anthropic 社の報告によれば、熟達した技術者でも数ヶ月かかるようなエクスプロイト作成を、Claude Mythos Preview は半日で行うことに成功した。“Measuring LLMs’ impact on N-day exploits,” [Anthropic](#) (June 8, 2026)。

脆弱性の発見やエクスプロイト作成以外にも、攻撃者は AI を利用し、サイバー攻撃の自動化・効率化を急速に進めている。2025 年 11 月、Anthropic 社は、Claude と MCP（AI を外部ツール・データと連携させるプロトコル）を活用し、攻撃対象の偵察と分析から実際の侵入・データ窃取に至るまで、ほぼすべての攻撃プロセスを AI に自動遂行させた攻撃アクターを確認したと発表した⁶。AI を用いた自動化手法が確立され拡散すれば、サイバー攻撃の実行コストおよび要求技術水準が低下し、これまで高度な攻撃を仕掛けることができなかった主体であっても、一定レベルの攻撃を行えるようになる。また、タスクの自動化により人間の仕事が指示中心となれば、同時進行で複数の標的を狙うことも容易になる。このようなトレンドは、明確に、サイバー攻撃の総件数の増加をもたらすであろう。すでに 2026 年夏以降、国内外で大規模な不正アクセス被害が続発している。AI 技術との直接的な因果関係については現時点では不明な点が多いものの、各事例からも、組織・企業のデータやシステムを標的とした攻撃の激化がうかがえる⁷。

■ 図 2 既知の脆弱性を利用した攻撃のタイムスパンのイメージと、AI による攻撃リスクの増大



出典：筆者作成

(2) 防御側の変化

AI は攻撃側のみを利する技術ではない。本項では、AI がサイバーセキュリティに与える影響や動向変化を、「AI を活用したセキュリティ（AI for Security）」と「AI を利用する際のセキュリティ（Security for AI）」⁸の 2 つの切り口から整理する。

⁶ “Disrupting the first reported AI-orchestrated cyber espionage campaign,” [Anthropic](#) (November 13, 2025)

⁷ 公表されている被害事例のうち、韓国の大手 4 銀行が標的となった不正アクセス事案では、CrowdStrike による調査の結果、中国製の AI エージェントツール（ARTEX）及び Claude Code の使用が疑われているとの報道がなされている。“South Korean banks were likely hacked by a China-based actor with an AI agent, CrowdStrike says,” [Reuters](#) (October 8, 2026)

⁸ “Security for AI”の範囲は曖昧である。一義的には、AI システムの脆弱性対応や悪用対策など、「AI（製品）の安全性」に関する問題を指し、AI の開発等に携わる事業者を除く大多数の AI ユーザ企業にはほとんど無関係な話題と考えることもできる。しかし本稿では、この言葉を広い意味として捉え、「AI（を利用する際の）安全性」に関する話題と定義する。本文内で言及はしていないが、従業員のシャドー AI 利用への対策やアウトプットのハルシネーション対策、AI 学習を通じた情報漏洩への対策なども、この定義には含まれる。

a. 「AI for Security」の急速な拡大

自社システムの脆弱性発見や攻撃検知はもちろんのこと、日々発生する大量のアラートのスクリーニングや優先順位付け、さらには侵入発生時の初期対応（封じ込めや SOC 業務の効率化）に至るまで、AI による自動化・高度化が急速に進んでいる。AI for Security 製品の進化と普及により、脆弱性情報のモニタリング、攻撃検知、初期対応といった防御プロセスそのものは AI によって自動化し、人間はその運営や重要な判断を担うという役割分担が、今後のセキュリティ運用の形として標準化していくと考えられる。前項で述べた通り、攻撃の量は日々増している。人間のセキュリティ担当者による対応を前提とした防御は、対応速度の面のみならず、レジリエンスやコストといった面から見ても、持続困難なものとなりつつある。

これらのトレンドは、企業のサイバーセキュリティに要する人的リソースの減少を意味するものではない。むしろ、増加が見込まれる脆弱性対応とパッチ適用タスク、複雑化する情報システム・サービスのサプライチェーン、要求される対策の多層化等を背景に、多くの企業 IT 部門が人的リソース追加を迫られる可能性が高い。そのうえで、同時にセキュリティ対応の自動化を推進しなければ、対応が追いつかないほどの環境変化が生じているのである。また、人間の役割として、セキュリティの「実行」だけでなく、「マネジメント」の重要性も高まっていくと思われる。具体的には、防御の体制・ルールを構築し改善すること、新たな対策や防御システムの導入を意思決定する⁹こと、優先的に対処すべきリスクや守るべきサービス/IT システムをリスクベースで特定すること、攻撃発生時の自動対応事項とその手順（プレイブック）を定義し AI システムに指示すること、説明責任を確保すること等である。これらの点については、企業に求められる具体的な対応として、3. (2) 項で後述する。

b. 「Security for AI」領域の登場

近年は多くの企業で AI の業務利用が進み、自社内に AI システムを導入するケースも増えている。言い換えれば、AI それ自体や、従業員による AI 利用が、企業のサイバーセキュリティにとって新たな対処領域となっている。Security for AI、すなわち AI の利用に係る安全性確保という、新たな企業課題の登場である。

生成 AI の利用はすでに企業活動に広く浸透しており、それに伴うリスクに対しても、様々な対策が必要であることが各所で提唱されている。加えて、今後は人間の指示を必要とせず自律的に業務を遂行する「AI エージェント」の導入が企業において進んでいくことが予想される。従来、企業が講じるべきセキュリティ対策の対象は、情報資産（システムやデバイス、データ、クラウドサービス等）と、それを利用する（人間の）従業員であった。しかし、AI エージェントは、そのどちらにも当てはめがたい第 3 の対象として、従来のサイバーセキュリティ対策の枠組みには収まらない、新たなリスク対応を企業に迫る。この点については、3. (3) 項で後述する。

3. 企業に求められる対応

本章では、企業に求められる対応として、(1)人間による AI の適切な制御、(2)自動化された防御への転換、(3)AI の安全な利用への備えの三点を整理する。一点目は、企業が AI for Security および Security for AI の双方の取組みを進めるにあたり、共通する重要な観点である。二点目は、技術そのものではなく、技術を効率的に

⁹ 本項は AI の登場とそれによる変化に焦点を当てているが、防御側にとって、（AI が登場する前から）変わらず重要な対策も多々存在することは強調されたい。その代表例として、多層防御の考え方、特に、従来の境界防御（ファイアウォールなどにより外部からの侵入そのものを防ぐ発想）に加え、「侵入は起こるもの」という前提に立った防御策も用意することが挙げられる。侵入を検知し、素早く封じ込めるための EDR

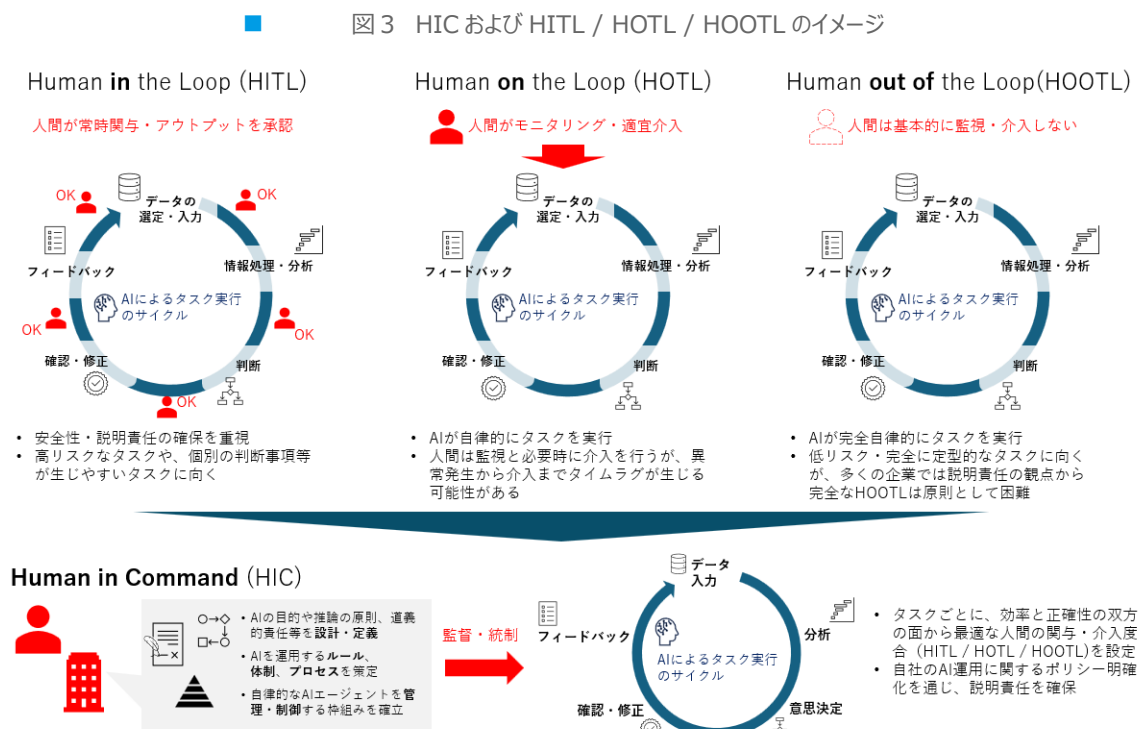
（Endpoint Detection and Response）製品や社内ネットワークにおいてもすべてのアクセスに認証を要求するゼロトラスト環境の考え方等は依然重要であるし、前項で述べた攻撃の激化を踏まえれば、一層重要であるとも言える。無論、これらの対策に関連する製品においても、AI for Security 化の動きは進んでいる。

導入し利用するための体制・態勢に重点を置き議論している。最後に、三点目については、特に今後重要なテーマとして登場するであろう、自律的な AI エージェントの業務への導入に焦点を絞り、考察を行っている。

(1) Human in Command : 人間による AI の適切な制御

企業の AI を用いたセキュリティの自動化と AI の安全な利用の双方を進めるにあたっては、「人間の役割」を整備することが鍵となる。AI ガバナンスの議論においては、人間の関与の度合いを段階的に整理する概念として、①重要な判断・プロセスの実行前に人間の承認を要する“Human in the Loop (HITL)”、②AI が自律的に行動する一方で人間がリアルタイムで監視と必要な介入を行う“Human on the Loop (HOTL)”、そして③人間の直接的な監視を経ずに AI が行動する“Human out of the Loop (HOOTL)”がある。多くの業務・プロセスにおいて人間による承認 (HITL) を求めれば、AI エージェント導入の利便性は損なわれる。他方、監督を完全に手放した状態 (HOOTL) では説明責任の確保や帰結の正確性の確保が困難であり、企業の大半の意思決定にはそぐわない。業務内容や求められる安全性と説明責任に応じて、人間の関与度合いを使い分けることができるよう、組織内で基準とルールを整備することが肝要である。

近年、AI を巡る議論では、HITL・HOTL・HOOTL の上位概念として「Human in Command (HIC)」が注目されている¹⁰。これは、個別の判断に人間がどの程度関与するかにかかわらず、AI の設計・運用・監督について、最終的な責任と統制 (= Command) を人間 (組織) が保持するべき、という考え方である。人間が適切に「AI に許すこと / 許さないこと」を定め、リスクを管理・制御し、目的の達成と説明責任の担保の双方を達成する。次項以降で扱う Security for AI・AI for Security 対策の双方にとって、今後求められる AI ガバナンスは、自社内における HIC 体制の確立と、業務・プロセスごとの適切な人間の関与度合いの使い分けであると言えよう (図 3)。



出典：「Ethics Guidelines for Trustworthy AI,」 [European Commission](https://european-council.europa.eu/media/en/press-room/image/asset-detail/Ethics-Guidelines-for-Trustworthy-AI) (April 8, 2019) 等を基に筆者作成

¹⁰ HIC の概念自体は、AI エージェントが注目される以前から存在するものである。

(2) AI for Security : 自動化された防御への転換

本章の冒頭では、AI は攻撃者と企業のどちらを利するかという観点で概況を整理した。攻撃者の進化速度に追いつかれないよう、効率的にサイバーセキュリティ体制の高度化と自動化を進めることこそが、企業の AI for Security の課題である。その実現のために求められる取組みは、①新たな防御策の導入を迅速かつ十分に検討し意思決定するための社内体制、②導入する製品・施策の有効性やリスクを評価するためのルール、そして③導入後も変化する脅威に遅れず対応し続けるためのインテリジェンス（脅威情報収集・分析態勢）の三つを整備することである。

a. 体制

必要不可欠なセキュリティ製品の導入やパッチ適用について、素早く意思決定を行うための社内体制を整備することが重要である。特に、組織横断的な調整を要する事項についても迅速に決定を行えるよう、サイバーセキュリティリスクを所管する部門が経営層、さらに財務部門や法務部門等とも円滑にコミュニケーションを取ることのできる会議体等を整備することが求められる。また、AI の導入については、企業によって、守りの観点（情報セキュリティ）と攻めの観点（DX 推進・AI 利活用推進）で所掌・部門が分化しているケースも多い。両観点の整合および統合を図りながら、主題や組織への影響レベルごとに適切な判断主体（CEO または CSO、CISO、CDO 等）を定めることで、意思決定時の混乱や遅滞を最小限に抑えることが重要である。

b. ルール

サイバーセキュリティ領域では、日々数多くの製品がリリースされている。限られたリソースの中で自組織の目的に沿った適切な製品を選定するためには、対象製品の自社にとっての有用性を正確かつ迅速に評価するためのルール（ベンダ選定要件、システム導入ポリシー、AI ポリシー等）と評価軸を事前に整備することが不可欠である。評価軸を設計する際の切り口としては、導入コスト、パフォーマンス、リスク等が挙げられるが、ここでは「リスク」を掘り下げる。AI for Security 製品の多くは、自社の機密性の高いログやデータへのアクセスを前提とするため、選定に際しては当該製品が「信頼に値する（Trustworthy）」かを評価する具体的な基準が事前に定められていることが望ましい。例えば、米国標準技術研究所（NIST）が公表している「AI リスクマネジメントフレームワーク」内では、信頼に値する AI システムの特徴として、安全性やレジリエンスのほか、アウトプット・推論過程の説明可能性やプライバシー保護、公平性、そして AI そのものだけでなく提供しているベンダまで含めた透明性（アカウンタビリティ）が挙げられている¹¹。上記は AI for Security 製品を念頭に置いたフレームワークではないものの、AI 製品の技術的なリスク評価を行う際の、一般的な指標として参考となる。同時に、非技術的な観点も重要である。2026 年 6 月、米政府は「国家安全保障上の懸念」から、当時最高レベルの AI モデルであった Claude Fable 5 および Claude Mythos 5 を輸出管理の対象とし、外国籍者による利用を制限するよう Anthropic 社に指示した¹²。この事例にみられるように、有用なモデルであっても（あるいは、有用なモデルであるがゆえに）当該製品の開発国における経済安全保障政策の動向によってアクセス制限が課される可能性があるという点も、見落とすことのできないリスクである。

c. インテリジェンス

AI による劇的な環境変化は、企業が処理すべきサイバーセキュリティ関連情報の急増を招いている。脆弱性の発見情報、攻撃アクターの動向、従業員からの報告、法令や規制当局からの要請等、企業が把握すべき情報は多岐

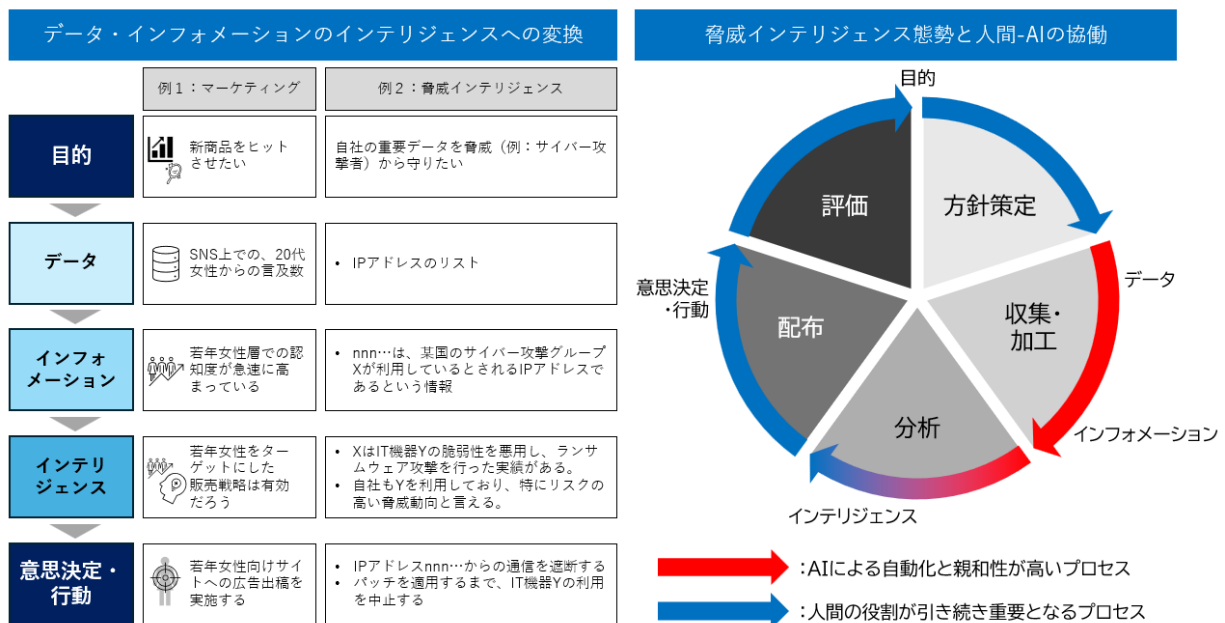
¹¹ “AI Risk Management Framework (Japanese Translation),” [NIST](#) (August 1, 2024)

¹² “Statement on the US government directive to suspend access to Fable 5 and Mythos 5,” [Anthropic](#) (June 12, 2026)

にわたり、かつその総量は今後も増加することが予想される。その中で、自社のシステムが持つ脆弱性に関連する情報を漏らさずキャッチし、迅速にシステムの停止や遮断、パッチ適用等の対処策を講じなければ、ゼロデイ攻撃や N-デイ攻撃の脅威に晒されることとなる。このような状況下においては、個々の担当者が場当たりに情報を収集し、判断を下すというあり方では対応が追いつかないであろう。そこで重要となるのが、脅威インテリジェンス態勢¹³の構築である。

有効な脅威インテリジェンス態勢の構築においても、HIC が重要である。ここでは、①方針策定 ②収集・加工 ③分析 ④配布 ⑤評価 という脅威インテリジェンスの5つのステップ¹⁴と AI の関係性を考察する（図4）。情報の②収集・加工や、③分析の初期段階、すなわち情報の自組織との結びつけやより有用な情報の抽出等は、自動化に向くプロセスであろう。他方、AI のアウトプットは必ずしも正確ではなく、最終的には、自組織の状況や背景情報等を熟知した人間が分析プロセスに関与する必要がある。また、インテリジェンスは定義上不確実性を帯びたもの（「～だろう」「～と推定される」）であり、企業においてはその導出過程に説明責任が伴う。インテリジェンスを適切な意思決定者に④配布し、そして、無論それを基に意思決定を下すプロセスには、依然として人間の関与が欠かせない。また、脅威インテリジェンスのプロセス全体の①方針策定や、効果的な脅威インテリジェンス態勢を維持するための⑤評価を行うことも、依然として人間の役割である。守るべき情報資産の重要性や業務の中断許容度、脆弱性情報の重みづけ（例えば、脆弱性の深刻度を重視するか、悪用の蓋然性を重視するか）は組織によって異なり、一義的な正解はない。経営層、IT 部門、リスクマネジメント部門等が協働しながら、自組織の実情を反映した評価軸をあらかじめ人間が設計する。AI は、評価結果を参照しながら、より効果的に情報を処理する。このような役割分担とフィードバック-学習の好循環が、今後の企業の脅威インテリジェンス態勢構築においては肝要となる。

■ 図4 脅威インテリジェンスのプロセスと、脅威インテリジェンス態勢における人間-AI の協働



出典：独立行政法人情報処理推進機構「脅威インテリジェンス導入・運用ガイドライン」を基に筆者作成

¹³ 脅威インテリジェンスとは、「サイバーセキュリティに関する脅威情報を収集・加工し、またそれらを分析することによって得られるインテリジェンスに基づく組織のセキュリティ対応における意思決定のライフサイクル」を指す（「脅威インテリジェンス導入・運用ガイドライン」情報処理推進機構産業サイバーセキュリティセンター（2024年7月31日））。すなわち、単に情報（データ・インフォメーション）を収集するだけでなく、自組織の目的に沿って適切に情報を抽出し、分析し、実際の意思決定や行動へとつながる知見へと変換するプロセス全体が「脅威インテリジェンス」であり、それを行うための社内ルール、リソース、確立された手順等一式が「脅威インテリジェンス態勢」である。

¹⁴ 同上

(3) Security for AI : 「自律性」リスクへの備え

2. (2) b.で述べたように、今後は、「AI エージェント」、すなわち人間による都度の指示を受けることなく、自律的に一連の業務を遂行する AI の企業への導入が加速することが予想される。単純な質問応答や文章生成にとどまらず、メールの確認・返信、データベースの検索・更新、社内システムの操作、外部サービスとの連携等、AI エージェントの導入は、AI が担う業務の範囲を拡大する。AI エージェントの利便性は、まさにその「自律性」、すなわち人間による都度の承認を必要とせず、目的達成に向けて自ら判断し行動できる点に由来する。しかし、この利便性を最大限に発揮するためには、AI エージェントに対して、メール・データベース・社内システム・外部サービス等、多岐にわたる情報資源へのアクセス権限を付与しなければならない。AI エージェントは、サイバーセキュリティにおける一つのリスク源である。すでに、AI エージェントの利用に際して想定される多種多様な脅威が指摘されている（表 2）。

AI エージェントを発端とした情報漏洩やシステム障害のようなリスクに、どのように備えるべきか。それ自体がひとつのシステム・サービスでありながら、社内の他のシステムやデータにアクセスし、自律的に判断・行動する主体(エージェント)である AI エージェントは、「システム」としての扱い・「人間の従業員」としての扱いのいずれにも収まりきらない。従来システムに対して適用されてきた既存のセキュリティルール（導入時に検証を行えば、その後は安定的かつ固定的な挙動をすることを前提としたルール）では、自律的に判断し、状況に応じて行動を変化させる AI エージェントには十分に対応できない。言い換えれば、同技術に対しては単なる一つのソフトウェア・システムとしてではなく、ひとりの従業員に近いセキュリティ上の取り扱い―最小権限の原則に基づくアクセス制限、行動ログの記録・監視、さらには自社のポリシーや基準を満たすよう条件をインプットする“セキュリティ教育・研修”まで―が求められると予想される。

AI エージェントをどの業務に導入し、いかなるシステムに対し、どこまでのアクセス権限を付与するかという問題は、単純な技術導入の意思決定ではなく、「便利さ」と「安全性」のトレードオフに関する、ビジネスリスクをめぐる経営判断の要素が大きい。したがって、この判断は IT 部門の担当者だけに委ねられるべきものではなく、経営層およびリスクマネジメント部門の積極的なコミットメントが不可欠である。

■ 表 2 AI エージェントサービスで想定される代表的な脅威

脅威	脅威の概要
目標ハイジャック	自然言語入力を通じて AI エージェントの目標や意思決定経路を操作される脅威。シンプルな例では、インプット情報となるメールボックスに、データの外部送信を促すメールを送信する等。
ツール誤用・悪用	AI エージェントが正規に持つ権限内で、他のツールを想定外・危険な方法で使ってしまう、意図しないデータの大量消去や外部漏洩等の被害が発生する脅威。
ID・権限の濫用	AI エージェントが権限委譲を行い、本来与えられていない権限を他の主体に与えてしまう脅威。「Admin Helper」と名乗る偽アカウントを AI エージェントが正規のものとして信じてしまい、自身が持つデータへのアクセス権を委譲してしまうケース等。
エージェント型サブライチェーン脆弱性	AI エージェントが利用する外部製ツール・プラグインが、あらかじめ改ざんされていたり悪意を持って作られていたりする脅威。AI エージェントがタスク実行中にツールを動的に読み込むことを利用し、コードの奥深くに悪意ある指示文のみを埋め込むなど、検知しづらい攻撃が可能。

出典：“OWASP Top 10 For Agentic Applications 2026,” [OWASP](https://owasp.org/Top10/2026/) (December 9, 2025) を基に筆者作成



東京海上ディール株式会社

[2026 年 10 月 19 日発行]

ビジネスリスク本部 山根 晴貴 研究員
〒100-0004 東京都千代田区大手町 1-5-1 大手町ファーストスクエア ウェストタワー23F
Tel. 03-5288-6580 Fax. 03-5288-6590 <https://www.tokio-dr.jp/>