World Trend Foresight

フロンティア AI を正しく恐れよ

—サイバーセキュリティへの影響を捉え直す—

2026 年 7 月

はじめに

新しい能力が登場したとき、その受け止め方は立場によって大きく食い違うことがある。フロンティア AI モデル[1]のサイバーセキュリティへの影響に関する議論も、その一例である。「サイバーセキュリティが根本から覆る」という危機を強調する声がある一方で、「宣伝色が強いのではないか」という見方もある。期待と不安が、同時にふくらんでいる。

本稿は、極論には与しない。先に結論を述べれば、フロンティア AI の影響は無視できないものの、現時点でサイバーセキュリティの原則を変えるものではない。むしろ、「自社が何を持っているかを把握し、自社に適したリスク対応策を選択し、継続する」といった基本の徹底が、いっそう重要になる。

本稿が主に想定する読者は、ユーザー企業においてサイバーセキュリティを専門としない経営層、および事業部門のビジネスパーソンである[2]。誇張にも過小評価にも傾かず、技術・実務・資源の三つの側面から、フロンティア AI によってサイバーセキュリティの「何が変わり、何が変わらないのか」を示す。第 1 章では技術の観点からサイバー攻撃能力の評価を、第 2 章では実務の観点から脆弱性対応のボトルネックを、第 3 章では資源の観点からフロンティア AI 導入のハードルと経営に求められる構えを論じる。**図表 1** は、本稿の見立てをまとめたものである。

図表 1　フロンティア AI によるサイバーセキュリティへの影響：「変わること」と「変わらないこと」

観点	変わること	変わらないこと
技術	**スピード** （AI が攻防双方の作業効率を底上げする）	**構造** （攻撃側は弱点を突き、防御側は全体を守る）
実務	**量** （脆弱性発見の件数が増える）	**ボトルネック** （影響評価・テスト・変更管理・適用判断が難所）
資源	**効率** （自社に最適な「人間−AI」の役割分担を探る）	**制約** （利用コスト、利用条件、人的資源の制約は残る）

（出所）筆者作成

[1] 既存モデルの能力を超える最先端の汎用 AI モデルであり、社会に対し深刻かつ新たなリスクをもたらす可能性があるもの。

[2] 本稿では主にシステムを利用する企業の視点から論じている。ソフトウェアを開発・提供する企業においては、自社製品に含まれる脆弱性の把握やパッチ提供が求められる点で、異なる対応が必要となる。

第 1 章　＜技術＞ フロンティア AI によるサイバー攻撃能力の評価

フロンティア AI のサイバー能力向上は、2025 年からの大きな潮流である。2025 年半ば、Google は AI エージェント「Big Sheep」が悪用寸前の脆弱性を攻撃者に先んじて発見・修正した事例を公表した[3]。2026 年に入ると業界の動きは加速する。同年 2 月、OpenAI は自社モデルを社内基準で初めて「サイバー能力が高い（high capability）」段階であると認定し[4]、3 月には脆弱性を自力で探す AI を公開した[5]。同モデルは短期間に多数の欠陥を発見し、複数の正式な脆弱性番号（CVE[6]）の登録に至ったと報告されている。4 月には、Anthropic がフロンティアモデル「Claude Mythos Preview」を公表する[7]。このモデルが持つサイバー攻撃能力が、メディア等で広く注目を浴びたことは記憶に新しい。同社は、このサイバー攻撃能力は攻撃用に訓練したものではなく、文章やプログラムを書く力・考える力・自分で作業を進める力といった汎用的な能力が伸びた副産物だと説明している。そのうえで、悪用を懸念して一般には公開せず、「Project Glasswing」を通じて総額 1 億ドル分の利用クレジットの提供をコミットし、テクノロジー企業やセキュリティ企業と連携しながらソフトウェアの脆弱性発見・修正を支援した[8]。5 月にはその途中経過として、脆弱性発見から修正までの貢献が報告されている[9]。OpenAI も同月、本人確認を経た防御者のみに同社モデルの特定バージョンを提供する枠組みである「Daybreak」を整備している。このように、各社そろって強力なサイバー能力を持つモデルを開発し、それを限定的に提供するという手順を踏んでいる。

フロンティア AI の流動性は、6 月に入って一段と際立っている。Anthropic は 6 月 9 日に、後継モデル「Fable 5／Mythos 5」を公表したが、その 3 日後となる 6 月 12 日には米国政府が国家安全保障上の権限を根拠とする輸出管理指令を発出し、同モデルへの外国人によるアクセスを全面的に停止する措置が講じられた[10]。Anthropic は当該措置に対し、公表文書の中で、限定的な安全性上の懸念のみを理由に商用モデルの提供停止を求めることには反対の立場を示しており、業界全体の研究開発や提供に萎縮効果をもたらす可能性があると主張した[11]。その後、Anthropic は 6 月 30 日に Fable 5 の再展開と Mythos 5 へのアクセス再開を公表した。ただし、7 月 1 日から世界的に利用可能となったのは Fable 5 であり、Mythos 5 については一部の承認済みパートナー組織に限ってアクセスが再開された[12]。並行して、6 月 26 日には OpenAI も新モデル「GPT-5.6」を、政府の要請を受けて少数のパートナーに限るかたちで提供を開始した後[13]、7 月 9 日には同モデルを一般公開している[14]。わずか 1 か月あまりの期間で、2 社のフロンティア AI が相次いで政策判断の影響を受けたことになる。これらの事例は、フロンティア AI の利用可能性そのものが技術的成熟度だけでなく、政策判断によって広範に制約され得るという現実である。これらの動きは、フロンティア AI をめぐる環境の流動性の高さを端的に物語っている。

ここでの論点は、フロンティア AI のサイバー攻撃能力をいかに評価するかである。英国 AI セキュリティ研究所（AISI：AI Security Institute）は独立した立場で Claude Mythos Preview を検証し、サイバーセキュリティ専門

[3] Google (2025) “A summer of security: empowering cyber defenders with AI” https://blog.google/innovation-and-ai/technology/safety-security/cybersecurity-updates-summer-2025/

[4] OpenAI (2026) “GPT-5.3-Codex System Card” https://openai.com/ja-JP/index/gpt-5-3-codex-system-card/

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

[6] 「Common Vulnerabilities and Exposures」の略。ソフトウェアやハードウェアに存在するセキュリティ上の脆弱性（Vulnerability）や問題（Exposure）を一意に識別・管理するための共通の識別子（ID）を提供する仕組みおよびリストのこと。

[7] Anthropic (2026) “Assessing Claude Mythos Preview’s cybersecurity capabilities” https://red.anthropic.com/2026/mythos-preview/

[8] Anthropic (2026) “Project Glasswing: Securing critical software for the AI era” https://www.anthropic.com/glasswing

[9] Anthropic (2026) “Project Glasswing: An initial update” https://www.anthropic.com/research/glasswing-initial-update

[10] Anthropic (2026) “Statement on the US government directive to suspend access to Fable 5 and Mythos 5” https://www.anthropic.com/news/fable-mythos-access

[11] Anthropic (2026) “Statement on the US government directive to suspend access to Fable 5 and Mythos 5” https://www.anthropic.com/news/fable-mythos-access

[12] Anthropic (2026) “Redeploying Fable 5” https://www.anthropic.com/news/redeploying-fable-5

[13] OpenAI (2026) “Previewing GPT-5.6 Sol: a next-generation model” https://openai.com/index/previewing-gpt-5-6-sol/

[14] OpenAI (2026) “GPT-5.6 目標に合わせて拡張するフロンティアインテリジェンス” https://openai.com/ja-JP/index/gpt-5-6/

家向けの演習課題において 73％の成功率を記録したと報告している[15]。ただし同レポートは、この数字は「実際にはしばしば存在するセキュリティ機能が欠けている」条件で得られたものであり、実世界でのシステムを実際に攻略できるかは判断できないと明記している。Anthropic が公表したシステムカード[16]にも、「我々の全体的な結論は、壊滅的リスクは依然として低いということである」という記載がある[17]。いずれの主張も、新モデルの能力向上は明らかであるが、当該モデルによるサイバー攻撃の実現性に関しては、条件を慎重に見極める必要があるという主張は共通している。フロンティア AI がサイバー攻撃における人間のパフォーマンスに与える影響を評価した最新研究においても、類似した結論が示されている[18]。また、Google 脅威分析チームは、攻撃者が AI を使い始めていることを事実としつつ、「状況を根本的に変える画期的な能力を達成したことはまだ確認していない」と述べている[19]。総じて現時点では、フロンティア AI が攻撃的なサイバー能力を劇的に向上させるという証拠は限定的であるが、初心者の参入障壁を下げたり、初級レベルのタスクを自動化したりできる、という評価が妥当であると考えられる。

もっとも、攻撃側への影響を軽く見てよいわけではない。公開情報をかき集めて要点をまとめる、攻撃の段取りを作る、ソフトウェアの脆弱性を自動で探すといった作業は、AI によって確実に速くなっている。AI に作業を分担させる工夫により模擬的なペネトレーションテスト[20]をやり遂げる割合が高まったとの報告や[21]、複数の AI を連携させることで脆弱性を突く攻撃タスクの効率が、それまでの手法に比べ最大 4.3 倍に高まったという報告もある[22]。一般論として、効率が上がれば、同じ時間で試せる攻撃の回数も増える。一つひとつの成功率が低くても、試行が桁違いに増えれば、守りの薄い箇所が突かれる機会は確実に増えていく[23]。

結論として、フロンティア AI は、攻撃と防御を根本から別物に変えるものではなく、もとからある力を底上げする「増幅（uplift）」の機能を果たすものである。これは、今までできていたことの「量」と「スピード」が格段に増すことを意味する。そして、その能力はデュアルユース、すなわち攻撃にも防御にも適用できる。Anthropic が、危険な能力をまず防御側へ届けようとしたのも、この両面性をふまえてのことである。技術の面で変わったのは双方の作業の速さであり、変わらないのは「攻撃側は弱点を突き、防御側は全体を守る」という攻防の基本構造である。だからこそ、底上げされた攻撃に対して、自社の守りをどう強化するかが問われる。

第 2 章　＜実務＞ フロンティア AI は「基本の徹底」を強いる

第 1 章で見たとおり、AI は攻撃と防御の双方でタスクの量と処理スピードを押し上げる。報道でも言及されているとおり、見つかる脆弱性の数が増えることが、実務へのわかりやすい影響である。では、脆弱性の量が増えたとき、企業の現場では何が起き、どこに労力を割く必要があるのだろうか。

---

[15] UK AISI (2026) “Our evaluation of Claude Mythos Preview’s cyber capabilities” https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities

[16] AI モデルや AI システムに関する透明性を高めるために作成されるドキュメント。主にシステム概要、安全性評価、能力と限界、訓練データ、リスク軽減策などの情報が含まれる。

[17] Anthropic (2026) “System Card: Claude Mythos Preview” https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf

[18] RAND (2026) “Investigating the potential use of frontier AI models for offensive cyberattacks: A human uplift study” https://www.rand.org/pubs/research_reports/RRA3892-1.html

[19] Google (2026) “GTIG AI Threat Tracker: Distillation, Experimentation, and (Continued) Integration of AI for Adversarial Use” https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use

[20] システムを疑似的に攻撃して、脆弱性を洗い出す検査。

[21] Deng, G. et al. (2024) “PentestGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing” USENIX Security Symposium, https://www.usenix.org/system/files/usenixsecurity24-deng.pdf

[22] Zhu, Y. et al.”Teams of LLM Agents can Exploit Zero-Day Vulnerabilities” https://arxiv.org/pdf/2406.01637

[23] サイバー空間における機会論の考え方は以下を参照されたい：PwC Intelligence（2026）“犯罪学から見たサイバーセキュリティ Vol.1 ＜日常活動理論が導くサプライチェーンセキュリティの論理＞” https://www.pwc.com/jp/ja/services/consulting/intelligence/assets/pdf/world-trend-foresight-109.pdf

脆弱性への対応は、「見つけたらすぐ直す」という単純な作業ではない。まず情報を取得し、自社のどのシステムが該当するかを棚卸しし、危険度を判断し、いつどう直すかを決める。ここで注意したいのは、弱点の深刻さと、実際の攻められやすさは別だという点である。深刻度が高くても、それがそのまま狙われやすさを意味するわけではない。そこで、実際に攻撃へ使われているかどうかなどに応じて、対応を「通常の更新でよいもの」「早めに手を打つもの」「可能な限り速やかに対応するもの」へと仕分ける枠組みが示されている[24]。また、対応のしかたも一通りではない。国際的に広く参照されているパッチ適用ガイドラインでは、脆弱性を修正する以外にも、あえて受け入れる・影響を小さくする・その機能の利用をやめるといった複数の選択肢があると整理している[25]。修正するか、いつ修正するかといった判断は、自社の事情に応じて見極めるべき「文脈依存型」の判断なのである。

とりわけ難しいのは、今まさに稼働している本番環境への適用段階である。高リスクだからといって、ただちに本番環境へ修正を入れればよいとは限らない。業務停止やシステム障害を避けるためには、影響評価、テスト、変更管理、適用タイミングの調整といった工程が必要になる。とくに金融など、安定稼働と説明責任が重い業種では、この工程そのものが大きな制約となる。したがって、実務上の難所は「脆弱性を見つけること」よりも、むしろその先にある「安全に、どの順で、どこまで直すか」を決め、実行することにある。

しかも、脆弱性の報告件数は年々増え続けている。その結果、すべての脆弱性を同じように評価・対応すること自体が現実的ではなくなりつつある。実際、NISTも2026年に運用方針を見直し、影響の大きい脆弱性に重点を置くかたちへと転換している[26]。

同様の問題は、ユーザー側でも深刻である。脆弱性対応において優先順位付けが死活的に重要であることは、2023年時点でデータにより示されている。企業データを集めて分析した調査によれば、組織がひと月に処理できる脆弱性は、抱えている全体数のおよそ10%にすぎない[27]。修正にかかる時間も対象によって大きく異なり、OSなどの脆弱性が半分に減るには30日ほどで済む一方、ネットワーク機器ではおよそ1年を要するという結果もある[28]。つまり、脆弱性発見の量は急増しているが、危険につながるのは一部であり、しかもその一部すら直しきれていない。自社の対応能力を遥かに上回る脆弱性の量を前提とすることになるため、高リスクなものを正しく選び出す優先順位付けの精度がいっそう問われることになる。

では、防御側のプロセスのどこにAIが効き、どこに効きづらいのだろうか。外部情報の取得や自社システムの棚卸し、初期の仕分けは、AIによる自動化になじむだろう。ただし、機械的な予測だけでは、決定的な判断材料にならない。自社を取り巻くその時々の条件と併せての判断が必要になる。ある脆弱性が自社のどの業務にどう響くかという判断、いつ対応するかという決定、資産の重要性区分やパッチ適用の実行などは、人の手に残る。だからこそ、成果を分けるのは基本施策の完成度であると言える。例えば、自社資産の把握は防御の出発点である。「自分が把握していないものは守れない」という原則は、企業が最低限理解しておかないとならないものである。さらに、守る対象を絞り込む簡素化・標準化は、危険度の見極めと修正の負担を軽くする。AIが注目される局面でこそ、サイバーセキュリティの基本の徹底が生命線となる。

つまり、実務の面で変わるのは「見つかる脆弱性の量」であり、変わらないのは「対応の難所が発見の先にある」という点である。それを支えるのは、資産把握・パッチ運用・標準化といった基本施策の完成度にほかならない。もっとも、これらの工程をAIにどこまで任せられるのかは、費用や資源という別の制約とも絡む。次章では、資源の観点からフロンティアAIとの関わり方を掘り下げていく。

---

[24] CISA “Stakeholder-Specific Vulnerability Categorization (SSVC)” https://www.cisa.gov/stakeholder-specific-vulnerability-categorization-ssvc

[25] NIST (2022) “Guide to Enterprise Patch Management Planning: Preventive Maintenance for Technology” https://csrc.nist.gov/pubs/sp/800/40/r4/final

[26] NIST (2026) “NIST Updates NVD Operations to Address Record CVE Growth” https://www.nist.gov/news-events/news/2026/04/nist-updates-nvd-operations-address-record-cve-growth

[27] Cyentia Institute (2023) “The Pithy P2P: 5 years of vulnerability remediation & exploitation research” https://www.cyentia.com/pithy-p2p/

[28] Cyentia Institute (2023) “Patching, Fast and Slow” https://www.cyentia.com/patching-fast-and-slow/

第 3 章　＜資源＞ 利用コストと利用条件の制約を知る

本章では、フロンティア AI をサイバーセキュリティに広く適用する際の、資源の問題を考えてみたい。ユーザー企業にとっての主な制約は、利用コストと利用条件である。高性能な AI は有用であっても、常に安価かつ安定的に使えるとは限らない。これらの制約は、AI への全面的な依拠を難しくする。

当たり前のことであるが、AI 利用には費用が掛かる。脆弱性をしらみつぶしに探させる使い方は、AI に何度も重い推論をくり返させ、その計算量に応じた費用がかさむ。第 1 章で触れた Anthropic の探索では、一つのソフトウェアの脆弱性を調べるのに、2 万ドル弱（1ドル 160 円換算でおよそ 320 万円）を要したと報告されている[29]。これは一つの対象についての額である。多数のシステムを抱える企業が、同様の手法で脆弱性を調査する場合は、その費用対効果を十分に考慮する必要がある。

また、こうした利用条件は、ユーザー企業の社内事情だけで決まるものではない。高性能な計算資源は、少数のクラウド事業者と特定地域に集中したサプライチェーンの上に成り立っており、その提供条件は、設備制約や政策判断にも左右される。電力や計算資源の逼迫は、ユーザー企業にとって直接の問題というより、価格、応答性能、利用制限、利用停止といったかたちで間接的に跳ね返る制約である。第 1 章で見た特定 AI モデルへのアクセス停止のように、能力そのものが政策判断で突然絞られることもある。特定の基盤や特定のモデルに依存するほど、価格変動や供給途絶の影響を受けやすくなる。

ここから導かれるのは、人と AI の役割分担を前提にした運用である。任せてよいかどうかの分かれ目は、その作業が定型かどうかだけではない。結果に対して人が責任を負う必要があるか、さらに規制や内部統制の観点から人間の確認・承認が求められるかどうかが境目になる。情報の収集と整理、初期の仕分けの下書き、定例的な照合のように、人が責任を負わずに済む作業は AI に任せやすい。一方、業務への影響の判断や対応の最終決定のように、結果に責任がともなう領域は人が担うべきである。どこまでを AI に委ねられるかは、効率性だけで決まるわけでもない。高リスクな操作や業務影響の大きい変更については、規制、内部統制、説明責任の観点から、人間の確認や承認を組み込むことが求められる場合がある。したがって、実務上の論点は「どこまで自動化できるか」ではなく、「どの工程を自動化し、どの工程に人間関与を残すべきか」を設計することへ移っていく。一部をあえて人の手に残すのは、後退ではなく、現実に即した最適化と捉えるのがよい。

この構えは、抽象論に留まらず、いくつかの具体的な経営判断に落とし込める。第一に、システムの簡素化である。守る対象が複雑に入り組んでいるほど、危険度の見極めも修正も難しくなる。まず仕組みをシンプルにすることが、その後のすべての土台になる。第二に、データとプロセスの標準化である。扱う情報や仕事の進め方を揃えることで、対応のばらつきが減り、速さと品質が安定する。第三に、標準化を土台にした自動化の範囲拡大である。手順が揃った定型の工程ほど、AI に安心して任せやすい。そのため標準化がより進むと、AI に委ねられる領域は無理なく広がっていく。第四に、自社の事業に見合った人的資本の投入の見極めである。自動化できる範囲が広がっても、最終的な判断は人に残る。限られた人材を、自社の事業にとって最も重要な判断へどう振り向けるかが、運用の質を分ける。この四つは、図表 1 で「変わらない」とした、人に残る判断と基本対策の練度を、経営の手で着実に高めるための道筋にほかならない。

資源の面で変わるのは AI 導入を前提とした「人・プロセス・技術への予算分配」であり、変わらないのは「費用や資源の制約が続くこと」である。経営が問うべきは、AI を入れるか入れないかという二者択一ではない。AI の時代に、自社のセキュリティ運用をどの原則のうえに、どう築くか、という一点にある。

[29] Anthropic (2026) “System Card: Claude Mythos Preview” https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf

おわりに

本稿は、フロンティア AI がサイバーセキュリティにもたらす変化を、技術・実務・資源の三つの側面から見てきた。技術の面では、攻める側も守る側も速くなるが、両者の基本的な構図は変わらない。実務の面では、見つかる弱点の量は増えるが、対応の難所が「発見の先」にあるという事実は動かない。資源の面では、AI 利用コストと条件を考慮しながら、予算分配を再考する必要がある。

三つの側面を貫く結論は、一つに集約できる。フロンティア AI は、脅威を一新する魔法でも、すべてを肩代わりする万能薬でもない。その正体は、攻める側と守る側の双方の力を底上げする「増幅」であり、この増幅に最も効くのは、システムを簡素にし、データと手順を標準化するという、足もとの基本を固める力にほかならない。その土台があってはじめて、自動化を無理なく広げ、限られた人の力を本当に重要な判断へ振り向けることができる。ただし、増幅の度合いは、固定されたものではない。攻撃側の底上げが進むほど、基本に求められる完成度の水準も上がっていく。

だからこそ、経営が最後に向き合うべき問いは、「フロンティア AI を正しく恐れられているか」である。誇張と過小評価のあいだに立ち、増幅という現実に見合った備えを地道に積み上げること。それが、フロンティア AI の時代に組織が取りうる、最も確実な構えだと言える。

石原　陽平　マネージャー
PwC Intelligence
PwC コンサルティング合同会社

丸山 満彦　パートナー
PwC コンサルティング合同会社

村上 純一　パートナー
PwC コンサルティング合同会社

PwC Intelligence　統合知を提供するシンクタンク
https://www.pwc.com/jp/ja/services/consulting/intelligence.html

PwC コンサルティング合同会社
〒100-0004　東京都千代田区大手町 1-2-1 Otemachi One タワー　Tel：03-6257-0700