Research Report

第一ライフ資産運用経済研究所
Daiichi Life Group

2026 年 6 月 2 日

金融機関は AI 攻撃を AI で防げるのか

～Claude Mythos が突きつける「防御の速度競争」と金融インフラのレジリエンス～

政策調査部　主席研究員　柏村　祐

（要旨）

○　AI は、金融機関にとって業務効率化の強力な手段であると同時に、サイバー攻撃を高度化する道具にもなりうる。2026 年 5 月、金融庁は日本銀行と連名で、金融機関等に対して「フロンティア AI による脅威変化を踏まえた金融機関等の短期的な対応」を要請した。これは、AI サイバーリスクが将来の抽象的懸念ではなく、金融機関の経営・監督上の現在進行形の課題になったことを示している。

○　高度な AI モデル「Claude Mythos」は、ソフトウェアの脆弱性発見や攻撃経路の構築を加速しうるものとして、従来の脅威モデルに見直しを迫っている。もっとも、現時点で十分に防御された IT システムに対して攻撃を達成できると断定されているわけではない。重要なのは、AI を過度に恐れることではなく、攻撃側の探索速度と自動化が高まる可能性を前提に、防御側の対応速度を引き上げることである。

○　金融機関に問われているのは、「AI 攻撃を完全に防げるか」ではない。AI が未知の脆弱性や想定外の攻撃シナリオを生み出しうる世界で、銀行、証券、保険、決済、清算、資産運用を含む金融インフラ全体のレジリエンスを維持できるかである。そのためには、TLPT（脅威主導型ペネトレーションテスト）やレッドチーム演習だけでなく、重要サービス・IT システムの特定、技術負債の解消、リスクベースのパッチ適用、ベンダー契約の確認、多層防御、BCP、外部連携を一体で進める必要がある。

○　金融機関に求められるのは、AI を使うか使わないかという単純な選択ではない。攻撃側が AI によって探索と攻撃の速度を高める局面では、防御側も脆弱性の検知、ログ分析、異常検知、パッチ適用の優先順位付け、インシデント対応に AI を組み込み、対応速度を引き上げることが求められる。同時に、経営層、IT・サイバー部門、リスク管理部門、業務部門、監督当局、IT ベンダー、業界団体が連携し、金融システム全体で協調防衛体制を構築することが急務である。

1. Mythosが突きつけた「AIで攻撃される金融システム」の現実

2026 年 5 月 22 日、金融庁は日本銀行と連名で、国内の金融機関等に対し、「フロンティアAIに

よる脅威変化を踏まえた金融機関等の短期的な対応」を要請した。フロンティアAIとは、最先端の高度な推論能力や自律的な処理能力を備え、既存の業務やリスク管理の前提を変えうるAIを指す。この対応の背景には、AIの中でも高度な推論能力や自律的な処理能力を備えたフロンティアAIが、サイバー攻撃に悪用される脅威が急速に高まっているという危機感がある。こうした脅威は、海外の金融当局も重く見ている。スペイン銀行（Banco de España）が公表した『金融安定報告（Spring 2026)』では、米Anthropic社が開発したAIモデル「Claude Mythos」が名指しで取り上げられた。同報告書は、開発元であるAnthropicの公表情報に基づき、Mythosがソフトウェアの未知の脆弱性（ゼロデイ脆弱性）を特定し、それらを組み合わせて攻撃経路（exploit chain）を構築する能力を示したと整理している。一方で、同報告書は、Mythosの能力について独立した広範な検証はまだ十分ではないとも明記している。

この変化は、身近な例でいえば、泥棒がAIを使って町中の家の弱点を一瞬で探せるようになる状況に近い。そのとき、警備員が従来通り徒歩で見回るだけでは、守り切れない可能性が高い。金融機関も同じである。攻撃側がAIによって弱点の探索、攻撃手法の検証、侵入経路の構築を速めるなら、防御側も検知、封じ込め、復旧の速度を高めなければならない。問題は、金融機関が従来の人手中心の守り方だけで、この変化に対応できるかである。サイバー攻撃は、もはや一金融機関のIT障害にとどまらない。銀行、証券、保険、決済、清算、資産運用会社へと連なる現代の金融システムでは、売買執行の停止、清算・決済の遅延、基準価額算定の不能、換金対応の停止を通じて、市場の混乱や顧客の信頼低下に波及しうる。AIによってサイバー攻撃が速く・自動化される時代に、金融機関は従来の人手中心の守り方だけで、金融システム全体を守り切れるのか。本稿では、この問いを出発点に、金融機関と監督当局がAI時代の防御力をどのように高めるべきかを考察する。

2. AI時代の金融サイバー防衛の現在地

(1) Mythosは金融サイバー防衛の脅威モデルを変えた

AIの進化は、サイバー空間における攻防の前提を変えつつある。スペイン銀行のレポートに掲載された資料１は、Claude Mythosが情報システムにおける脆弱性の特定と悪用を加速・拡大させ、脅威モデルそのものを再構築しうる可能性を指摘している。Mythosが示した本質的な脅威は、AIがこれまで不可能だったサイバー攻撃を一気に可能にしたことではない。これまで高度なスキルを持つ専門のハッカー集団に依存していた「脆弱性の発見」「攻撃手法の検証」「複数手法の組み合わせ」という一連のプロセスの速度を、AIが大きく高めうる点にある。

一方で、同資料は「Project Glasswing」のような取り組みにも言及している。これは、広範な商業リリースの前に、一部の企業に限定的なアクセスを提供し、AIを活用したサイバー攻撃に対する防御戦略を開発させるためのプロジェクトである。金融機関にとって重要なのは、AIを単なる攻撃側の脅威として恐れるだけでなく、自らのシステムを守るための防御側の能力としても位置づけ直すことである。

資料 1　Claude Mythos AI モデル

	これは何ですか？	なぜ重要なのですか？
クロード・ミトス (Claude Mythos)	・ソフトウェアの脆弱性の特定と悪用における高度な能力を持つAIモデル ・米国の企業Anthropicによって開発	情報システムにおける脆弱性の特定と悪用を加速・規模拡大させ、脅威モデルを再形成する可能性がある
主な技術的リスク	・情報システムに対する攻撃の速度、規模、自動化の増加 ・その能力は現在、独立した広範な検証を待っているため、不確実である	・現在の防御は、サイバー攻撃の速度と新しいタイプによって圧倒される可能性がある ・既存のシステムでは処理が困難な、より同期された脆弱性悪用の波を生成する可能性がある
プロジェクト・グラススウィング (Project Glasswing)	より広範な商業リリースの可能性を前に、複数の米国企業にクロード・ミトスへの限定的なアクセスを提供し、防御戦略を開発させる	・参加企業に、この技術を活用したサイバー攻撃に対処するための防御戦略を開発する機会を与える
システム全体への影響	・経済のすべての分野におけるサイバーインシデントの増加 ・クロード・ミトスの能力は、他のモデルによって複製および強化される可能性がある	・すべての経済セクターにおけるオペレーショナル・リスクの増加

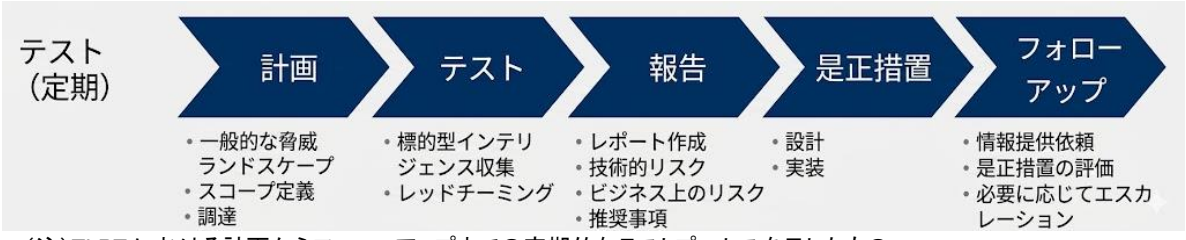
（注）Claude Mythos の概要と主要な技術的リスク、システム全体への影響を示したもの。
（出所）Banco de España "Financial Stability Report Spring 2026"より第一ライフ資産運用経済研究所作成

(2) TLPTは必要だが、それだけでは十分ではない

攻撃の速度と高度化が増すなか、防御のあり方も根本的な見直しが迫られている。IMF（国際通貨基金）のレポートに示された資料 2 は、脅威主導型ペネトレーションテスト（TLPT：Threat-Led Penetration Testing）のプロセスを示している。TLPTは、計画（Planning）、テスト実行（Testing）、報告（Reporting）、修復（Remediation）、フォローアップ（Follow-up）という一連の工程を通じて、実際の攻撃者の戦術・技術・手順を模擬し、組織の防御力を検証するアプローチである。

AI攻撃時代の金融サイバー防衛において問われるのは、最新のセキュリティ製品を導入しているか、立派な規程が存在するかではない。実際に高度な攻撃を受けたときに、それを迅速に検知し、被害を封じ込め、業務を復旧できるかという実効性である。特に、ひとたび停止すれば市場全体に影響を及ぼす金融インフラにおいては、売買執行や清算・決済を止めないためのレジリエンスが不可欠である。そのためには、机上のチェックリストによる点検から脱却し、TLPTやレッドチーム演習を通じた実戦型の検証を定期的に組み込むことが欠かせない。

資料 2　TLPT（脅威主導型ペネトレーションテスト）のフェーズ



（注）TLPT における計画からフォローアップまでの定期的なテストプロセスを示したもの。
（出所）IMF "Good Practices in Cyber Risk Regulation and Supervision" より第一ライフ資産運用経済研究所作成

ただし、Claude Mythosに象徴されるフロンティアAIリスクに対して、TLPTだけで十分だと考え

るのは危うい。TLPTは、既存の脅威インテリジェンスや想定シナリオに基づき、防御態勢を検証する仕組みである。もちろん高度な演習であるが、AIが未知の脆弱性を発見し、攻撃シナリオを自律的に組み合わせ、防御側の反応を学習する可能性を考えると、定期的なテストだけでは対応が後追いになりかねない。したがって、AI攻撃時代の金融サイバー防衛では、TLPTを「十分条件」ではなく「必要条件」と位置づける必要がある。金融庁・日本銀行の短期対応が示すように、金融機関に求められるのは、実戦型テストに加えて、重要サービス・ITシステムの特定、ソフトウェア構成やネットワーク構成の把握、技術負債の解消、パッチ適用のリソース確保、ベンダー契約の確認、リスクベースの優先順位付け、多層防御、BCP、外部連携を一体として整備することである。とくに重要なのは、脆弱性対応の速度である。フロンティアAIによって短期間に大量の脆弱性が発見され、修正プログラムが集中的に提供される場合、すべての脆弱性に同じ優先度で対応することは現実的ではない。金融機関は、自社の重要サービスに与える影響、攻撃が成立する蓋然性、外部公開システムかどうか、代替手段の有無を踏まえ、リスクベースで対応順序を決める必要がある。

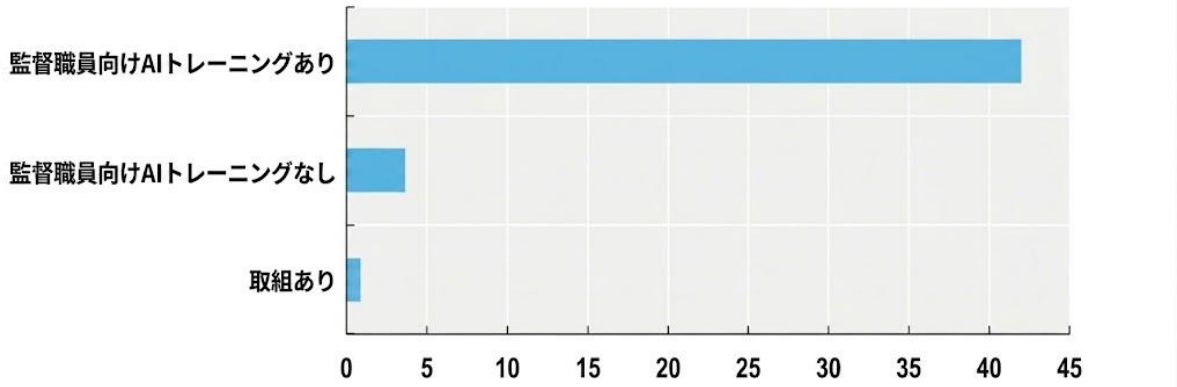
また、パッチを適用すればよいという単純な話でもない。パッチ適用に伴うシステム障害リスクと、パッチ未適用に伴うサイバー攻撃リスクを比較し、場合によってはテスト内容の合理的な縮小、WAFによる仮想パッチ、多要素認証、EDR、ネットワーク分離、特権ID管理などを組み合わせる必要がある。さらに、どうしても防御できない場合や、攻撃リスクが高まった場合には、重要サービスやITシステムを能動的に停止する判断基準もあらかじめ定めておかなければならない。つまり、AI攻撃時代の防御力とは、「攻撃を一度も受けない能力」ではない。攻撃を早く見つけ、影響を限定し、業務を継続または速やかに復旧する能力である。TLPTやレッドチーム演習は、その能力を検証する重要な手段である。しかし、それだけに依存すれば、過去の脅威シナリオに最適化された備えにとどまる可能性がある。金融機関は、テスト、日常的な脆弱性管理、パッチ運用、監視、BCP、外部連携を連続した防御プロセスとして再設計する必要がある。

(3) 監督当局もAI対応力を高める必要がある

金融システムの防衛は、個々の金融機関の努力だけで完結するものではない。OECD（経済協力開発機構）のレポートから引用した資料３を見ると、そのトレンドは明らかである。大多数のOECD諸国において、AIに関連する監督職員のスキルアップに向けた取り組みが進行中であることが示されている。同レポートによれば、金融分野におけるAIの監督には、従来の金融監督の専門知識に加えて、AIシステムの技術的理解や継続的な人材育成が必要不可欠であると指摘している。さらに、監督当局自身がAIを活用した監督テクノロジー（SupTech：Supervisory Technology）を導入することの重要性も強調している。

クラウドサービス、ITベンダー、決済ネットワーク、市場インフラが複雑に相互接続する現代の金融システムにおいて、AIを用いた攻撃の予兆は一企業の中だけで捉えきれない可能性がある。監督当局もまたAIの特性を深く理解し、SupTechを駆使して市場全体にまたがる異常を早期に把握・警告する能力を備える必要がある。

資料 3　OECD 諸国における監督職員向け AI スキル向上策の実施状況



（注）原資料では、3 つ目の項目が“Yes”と表記されている。具体的な取組内容は図中で明示されていないため、本図では「取組あり」と訳した。
（出所）OECD "Supervision of Artificial Intelligence in Finance" より第一ライフ資産運用経済研究所作成

3. 金融機関は「AIで守る」体制へ移行せよ

前節で確認したように、Mythos に代表されるフロンティア AI は、サイバー空間における攻防の速度を一段引き上げる可能性がある。金融機関側も、この変化を前提に防御体制を見直す必要がある。ただし、ここでいう「AI で守る」とは、AI に判断を丸投げすることではない。人間が検証・統制できる範囲を明確にしたうえで、AI を検知、分析、判断支援、復旧対応の補助線として組み込むことである。

金融機関が優先すべき対応は、次の 4 点に整理できる。

第一に、フロンティア AI への対応を経営課題として扱うことである。AI サイバーリスクは、IT・サイバーセキュリティ部門だけの問題ではない。重要サービスの停止、顧客対応の混乱、決済・清算の遅延、市場取引への影響、レピュテーション低下を通じて、金融機関の経営そのものに直結する。したがって、経営トップ、CIO、CISO、リスク管理部門、業務部門、財務部門が横断的に関与し、優先的に守るべきサービスと IT システムを特定する必要がある。守るべき対象が曖昧なままでは、脆弱性が大量に発見された局面で、限られた人員や予算をどこに集中すべきか判断できない。

第二に、防御プロセスの徹底的な高速化である。金融機関は、脆弱性の継続的な検知、ソフトウェア構成の把握、膨大なセキュリティログの分析、未知の脅威の異常検知、パッチ適用の優先順位付けに AI を組み込み、防御対応の時間を縮める必要がある。攻撃側が AI を用いて探索と攻撃の速度を上げるのであれば、防御側も AI を用いて検知と判断の速度を上げなければならない。とくに、外部公開システム、インターネットバンキング、取引システム、決済・清算関連システムなどは、リスクベースで優先順位を明確にし、パッチ適用や代替防御策を迅速に実行できる体制を整える必要がある。

第三に、実戦型テストを起点とした継続的な改善である。TLPT やレッドチーム演習は、AI 時代にも重要である。ただし、それらは単発の検査ではなく、防御プロセス全体を改善するための起

点として位置づけるべきである。演習によって明らかになった課題を、技術負債の解消、ログ監視の高度化、インシデント対応手順の見直し、パッチ運用、ベンダー契約、BCP に反映しなければならない。AI 攻撃時代には、テストを実施した事実だけでは十分ではない。演習で見つかった弱点を、運用改善、技術負債の解消、監視体制の見直し、復旧訓練に反映できるかが問われる。

第四に、金融インフラ全体での協調防衛体制の構築である。金融機関は、自社だけでサイバーリスクを完結的に管理できるわけではない。クラウド、IT ベンダー、共同センター、決済ネットワーク、市場インフラ、外部ソフトウェアに依存する以上、脅威情報の共有、緊急時の連絡体制、ベンダーの対応余力、SLA・SLO の確認、共同運営システムにおける責任分担が不可欠である。監督当局、金融機関、IT ベンダー、業界団体、金融 ISAC が連携し、AI を活用した監視、脅威情報のリアルタイム共有、早期警戒の仕組みを整える必要がある。

Mythos が投げかけた課題は、AI を使うか使わないかという単純な二分法ではない。攻撃側が AI を使って脆弱性探索や攻撃シナリオの構築を速めるのであれば、防御側もリスクを管理しながら、AI を防衛プロセスの一部として組み込まざるを得ない。攻撃側が AI によって速くなるのであれば、防御側も AI を活用し、検知、判断、封じ込め、復旧の速度を上げなければならない。金融システムの強靭性は、攻撃を完全に防ぐことだけで測られるものではない。攻撃をどれだけ早く見つけ、影響をどこまで限定し、顧客や市場への説明責任を果たしながら業務を復旧できるかが、AI 時代の金融サイバー防衛の核心になる。金融機関に求められるのは、「AI 攻撃を AI で完全に防ぐ」ことではない。AI が攻撃の速度を高める時代に、AI も活用しながら金融インフラ全体のレジリエンスを維持することである。

以　上

【参考文献】

・金融庁・日本銀行（2026）「フロンティア AI による脅威変化を踏まえた金融機関等の短期的な対応」
・Banco de España（2026）″Financial Stability Report Spring 2026″
・IMF（2026）″Good Practices in Cyber Risk Regulation and Supervision″
・OECD（2026）″Supervision of Artificial Intelligence in Finance″