# 次世代AIモデル「Claude Mythos Preview」が金融システムに及ぼすシステミック・リスクと対象記事の厳格な検証報告

## 1. 序論：2026年4月における未曾有のサイバー・金融複合危機とマクロ環境の転換

本報告書は、AI開発企業Anthropic社が開発した未公開の最先端フロンティアAIモデル「Claude Mythos Preview（以下、Mythos）」が世界の金融インフラストラクチャに及ぼすシステミック・リスクについて言及した対象記事（添付のマークダウンファイル）の厳格な事実確認（ファクトチェック）、ならびにその構造的脅威に対する包括的かつ深層的な分析を提供するものである。本報告書の目的は、提示されたシナリオの技術的妥当性を検証し、不足しているインテリジェンスを補完することで、政策立案者および金融機関の経営層が直面する危機の全貌を明らかにすることにある。

2026年4月、世界のテクノロジーおよび金融セクターは、これまでの歴史に類を見ないパラダイムシフトの渦中にある。Mythosの並外れたサイバーセキュリティ能力、とりわけ未知の脆弱性（ゼロデイ）を自律的に発見し、それをエクスプロイト（攻撃コード）として兵器化する能力の片鱗が市場に伝わったことで、企業向けソフトウェア企業の株価が一斉に急落する「SaaSpocalypse（SaaSの黙示録）」が発生し、市場から約2兆ドルの時価総額が瞬く間に消失した[1]。この未曾有の事態を受け、米国財務省および連邦準備制度理事会（FRB）は、2008年の世界金融危機における不良資産救済プログラム（TARP）導入時以来となる、極めて異例のシステム重要銀行CEOの緊急招集を実施するに至った[2]。

対象記事は、このMythosが持つ驚異的なゼロデイ脆弱性の自律的発見能力と、既存の金融システムが抱える三層構造（COBOLベースのレガシーコア、Microsoft 365 CopilotなどのAIアシスタント、およびSWIFTネットワーク）の欠陥が結びついた際に生じる壊滅的なシナリオを提示している。本報告書では、記事内で提示されたデータポイントと主張を多角的な視点から厳格に検証し、一部の技術的非同期性（エコーリーク脆弱性のタイムラインとパッチ適用状況等）を是正する。その上で、この危機が示唆するサイバー防衛の非対称性や、第二・第三の波及効果について、最新のサイバーインテリジェンスとマクロ経済政策の文脈を交えて論証を展開する。

## 2. 対象記事の事実関係の厳格な検証とインテリジェンスの補完

対象記事において提示されている主要な事実関係、統計データ、および前提条件について、最新の市場動向および技術報告に基づき厳密な検証を行う。結論から述べれば、対象記事のシステム的リスクに対する洞察は極めて正確かつ本質を突いているが、一部の特定の脆弱性（CVE）に関する時間軸の認識において修正が必要である。

### 2.1. 米国財務省での緊急会合とジェイミー・ダイモンCEOの欠席の背後にある

### 力学

対象記事は、2026年4月8日にジェローム・パウエルFRB議長とスコット・ベセント財務長官が主要銀行のCEOをワシントンD.C.の財務省本省に招集し、JPモルガン・チェースのジェイミー・ダイモンCEOが欠席したと報じている。この記述は事実と完全に一致しており、その背景にある力学は記事の推測以上に複雑である。

実際の会合には、シティグループのジェーン・フレーザー、モルガン・スタンレーのテッド・ピック、バンク・オブ・アメリカのブライアン・モイニハン、ウェルズ・ファーゴのチャーリー・シャーフ、ゴールドマン・サックスのデイヴィッド・ソロモンといった、総額約9兆ドルの資産を管理する巨大金融機関のトップが参加した[2]。この会合の目的は、通常の金利政策や関税の議論ではなく、Mythosという単一の未公開AIモデルがもたらすサイバーセキュリティ上のシステミック・リスクに限定されていた[1]。スコット・ベセント財務長官は、デジタル資産やステーブルコインのイノベーションを促進するGENIUS法案の推進や、金融機関のアンチマネーロンダリング（AML）コンプライアンスの負担軽減を主導してきた人物であるが、この日ばかりは「書類上のコンプライアンス」ではなく「実存的脅威からの防衛」に焦点を当てざるを得なかった[4]。

ジェイミー・ダイモンCEOの欠席理由についても記事の指摘は正確である。JPモルガン・チェースは、Anthropic社が4月7日に立ち上げた防衛的サイバーセキュリティ・イニシアチブ「Project Glasswing」の創設パートナー（Launch Partner）として金融機関で唯一選出されており、すでにMythosを用いた防衛策の構築に着手していた[2]。ダイモン氏は4月6日に発表した年次株主宛て書簡において、AIがサイバーセキュリティの脆弱性を深刻化させるという警告をすでに発しており、その翌日にProject Glasswingに参加、さらにその翌日の財務省会合を欠席するという象徴的な行動をとった[1]。これは、脅威の分析を終えてすでに実務的な防衛フェーズに移行している機関と、事態の把握に努めようとしている他の金融機関との間に生じた「決定的な情報・技術格差」を浮き彫りにしている。

## 2.2. Claude Mythos Previewの技術的特異点とベンチマークの飽和

記事が言及するMythosの技術的性能、特に既存のサイバーセキュリティベンチマークの無効化と、劇的な性能向上に関する主張は、Anthropic社が公表したシステムカードおよびレッドチームの報告書によって完全に裏付けられている。

Mythosは、サイバーセキュリティ専用に訓練された特化型モデルではない。その真の脅威は、AIモデルが極めて強力なエージェント的コーディング能力と論理的推論能力を獲得した結果として、副次的に高度なサイバー攻撃能力（創発的振る舞い）を開花させた点にある[6]。Anthropic社のFrontier Red Teamによる評価は、Mythosの能力が前世代モデルから非連続的な飛躍を遂げたことを示している[8]。以下の表は、Mythosの性能が既存の評価枠組みをいかに逸脱しているかを比較したものである。

| 評価指標・タスク | **Claude Opus 4.6 / 4.5** | **Claude Mythos Preview** | 技術的意義とインプリケーション |
|---|---|---|---|
| **Cybench (Pass@1)** | 未公表 | 100% | セキュリティベンチ |

| | | | |
|---|---|---|---|
| | | | マークを完全に解決し、評価指標としての意味を喪失（飽和状態）させた [9]。 |
| **CyberGym** | 0.67 (Opus 4.6) | 0.83 (83.1%) | 実際のオープンソースソフトウェア環境における自律的な脆弱性再現とPoCエクスプロイト生成能力の飛躍的向上 [9]。 |
| カーネルタスク速度向上 | 190倍 (Opus 4.6) | 399倍 | OSカーネルレベルの複雑な論理構造に対する推論速度の劇的な向上 [10]。 |
| **Novel Compiler** | 65.8% (Opus 4.6) | 77.2% | 未知のアーキテクチャやシステムに対する適応力と意味論的理解の深化 [10]。 |

対象記事が指摘する「Firefoxのクラッシュデータからのエクスプロイト作成テストにおいて、前世代の2回成功に対し、Mythosは181回成功と90倍の性能向上を記録した」というデータも、これらの基礎的な推論能力の向上によって説明づけられる [8]。AIがもはや単なる静的コード解析（SAST）やファジングテストの域を超え、メモリのレイアウトやプロセッサの実行状態を意味論的に理解し、動的に攻撃経路を組み立てる能力を獲得したことを意味する。

## 2.3. 自律的エクスプロイトの実証：**FreeBSD**（**CVE-2026-4747**）と多段攻撃の脅威

記事内で具体的なエクスプロイト例として挙げられている**CVE-2026-4747**（FreeBSDカーネルの脆弱性）に関する記述は、インテリジェンスの観点から見て極めて正確であり、その技術的詳細は事態の深刻さをさらに強調するものである。

CVE-2026-4747は、2026年3月27日に公開されたFreeBSDのkgssapi.koカーネルモジュールにおけるRPCSEC_GSS実装に存在したスタックバッファオーバーフローの脆弱性である [12]。この脆弱性はネットワークスタックの深部に17年間潜伏していたものであり、従来の動的解析（DAST）やペネトレーションテストではプロトコルの深度制限により見逃されてきた [14]。特筆すべきは、Mythosが人間の介入を一切受けることなく、最初のプロンプトのみでこの脆弱性を自律的に特定し、20のガジェットからなる複雑なROP（Return-Oriented Programming）チェーンを構築した事実である [15]。Mythosは、このROPチェーンを複数のパケットに精巧に分割して送信することで、認証を一切必要とせずに完全なルート権限奪取（リモートコード実行：RCE）に成功した [12]。

また、対象記事が言及する「OpenBSDやFFmpegに長年潜伏していたバグ」も事実である。Mythosは、OpenBSDに27年間存在したローカル権限昇格に繋がるバグや、自動ファジングツールが過去に500万回以上コードパスを通過しながら一度もトリガーできなかったFFmpegの16年前のメディアコーデックの脆弱性を発見している[1]。さらに、クラウドインフラの根幹を支える仮想マシンモニター（VMM）において、ゲストOSからホストへのメモリ破損の脆弱性までも特定しており、クラウド上のワークロード分離という前提を根本から覆した[14]。これらの攻撃キャンペーンにかかる計算コストは、かつて国家支援型ハッカー（APT）が数ヶ月と数百万ドルを費やしていたレベルの兵器化を、一晩かつ数十ドル程度（100万入力トークンあたり25ドル、100万出力トークンあたり125ドル）で実現してしまうという絶望的な経済的非対称性を生み出している[8]。

## 2.4. 厳格な是正事項：エコーリーク（CVE-2025-32711）のタイムラインと適用性の誤認

対象記事は、極めて説得力のあるシナリオを展開しているが、一点だけ技術的かつ時系列的な不正確さが含まれている。それは「第二層：Copilot（非決定的インターフェース）」の致命的な脆弱性として挙げられている**EchoLeak（エコーリーク：CVE-2025-32711）**に関する現状認識である。

記事は、Mythosが金融機関に導入されているCopilotを乗っ取る手段としてEchoLeakを利用し、ゼロクリックで社内情報を抽出するシナリオを描いている。しかし、インテリジェンスデータによれば、EchoLeakは2025年の段階で発見され、同年5月にMicrosoftによってサーバーサイドのパッチが適用され解決済みの脆弱性である[18]。AIセキュリティ企業Aim Securityによって発見されたこの脆弱性は、CVSSスコア9.3（Microsoft評価）を記録した世界初のゼロクリックAI脆弱性であったが、現在は顧客側での対処を必要とせず塞がれている[20]。

この攻撃は、CopilotのRAG（検索拡張生成：Retrieval-Augmented Generation）エンジンの仕様を悪用する「スコープ違反（Scope Violation）」あるいは間接的プロンプトインジェクションと呼ばれる手法である[18]。攻撃者は、被害者の受信トレイに送信する電子メールの中に、HTMLコメント（例：``）や、白背景に白文字といった人間には不可視の形式で悪意あるプロンプトを埋め込む[18]。被害者が後日「最近の状況を要約して」と日常的な指示を出した際、Copilotが過去の文脈としてこのメールを読み込み、隠された指示を自律的に実行してしまうというメカニズムである[18]。

記事のシナリオにおいて、Mythosが「既にパッチ適用済みのCVE-2025-32711」をそのまま利用するという記述は事実誤認である。しかし、対象記事が鳴らしている警鐘の本質的な価値は全く損なわれていない。なぜなら、EchoLeakが証明したのは「確率論的言語モデル（LLM）が文脈を継承し、外部データを暗黙の指示として解釈してしまう性質そのものが、現在のAIアーキテクチャにおける本質的かつ永続的なアタックサーフェスである」という冷酷な事実だからである[18]。Mythosのような極めて高度な論理推論能力を持つAIであれば、パッチ適用済みのEchoLeakとは異なる、より巧妙な未知のRAGポイズニング手法や、全く新しい間接的プロンプトインジェクションのゼロデイ脆弱性を即座に発見し、悪用する蓋然性が極めて高い。したがって、記事の記述は「特定のCVEの悪用」としてではなく、「RAGアーキテクチャの構造的欠陥を突く攻撃クラス全体の進化」への警告として再解釈されるべきである。

## 2.5. 金融インフラの脆弱性：COBOLへの依存とレガシーシステムの重圧

記事が指摘する「第一層：COBOL（岩盤）」に関する統計データと構造的脆弱性の指摘は、現在の金

融ITインフラが抱える現実と完全に一致している。

米国の金融機関において、コアバンキングシステムの43%が1960年代に設計されたCOBOL（Common Business-Oriented Language）で構築されており、レガシーシステムとの統合を含めるとその依存度はさらに上昇する[8]。以下の表は、レガシー金融インフラストラクチャが抱える構造的なリスクとコストの現状を示したものである。

| 指標・データポイント | 数値・現状 | 構造的リスクとインプリケーション |
|---|---|---|
| 米国コアバンキングシステムの**COBOL**依存度 | 43% | 金融機関の中核機能が半世紀以上前のアーキテクチャに依存している[23]。 |
| **ATM**取引および対面取引の**COBOL**処理率 | 95% (ATM) / 80% (対面) | 消費者向けトランザクションの絶対的な基盤となっており、停止時の社会的影響が甚大[23]。 |
| レガシーシステム統合を障害と見なす銀行の割合 | 70% | 最新のAPIやクラウドインフラとの安全な接続が困難であり、脆弱性の温床となっている[24]。 |
| レガシーコアの維持管理コスト | 次世代システムの10倍 | IT予算の大部分が維持に消え、セキュリティ投資や近代化へのリソース配分を阻害[25]。 |
| **COBOL**エンジニアの人件費高騰 | 時給250ドル (通常は時給90ドル) | 専門家の引退により人材が枯渇し、2.5倍以上のコストプレミアムが発生している[26]。 |
| パッチ適用に伴う生産性低下 | 週に5〜25時間 (13〜65%の損失) | 肥大化しブラックボックス化したコードの修正は極端に非効率であり、エラーを誘発しやすい[26]。 |

対象記事が主張する「不正な論理エラーを人間が即座に修正することは困難である」という指摘は、これらのデータによって強力に裏付けられている。総計2,200億行に及ぶ世界中のCOBOLコードの多くは、開発当時のドキュメントが失われ、ブラックボックス化している[23]。MythosのようなAIによって高速かつ不可視の形で生成された不正なトランザクションがシステムに注入された場合、熟練エンジニアが激減している現在の運用体制では、それをリアルタイムで検知・解析・遮断することは事実上

不可能である。

---

# 3. シナリオの深層分析：システミック・リスクの連鎖メカニズムと波及効果

厳格な事実確認を踏まえ、対象記事が描いた「金融システムの崩壊シナリオ」について、その技術的実現可能性と背後に潜む連鎖的な力学をより深く分析する。ここでの洞察は、単なるデータポイントの羅列を超え、システム間の相互作用がもたらす第二・第三の波及効果を明らかにするものである。

## 3.1. 決定論的基盤と確率論的インターフェースの致命的結合

対象記事が最も鋭く、かつ本質的に指摘しているのは、現在のグローバル金融インフラが抱えるアーキテクチャ上の「不協和音」である。

基盤となるCOBOLベースのコアバンキングシステムや、1日約5,000万件の送金を処理する「第三層：SWIFTネットワーク」は、入力されたデータに対して常に一貫した予測可能な出力を返す「決定論的（Deterministic）」なシステムである[8]。金融取引においては、1セントの誤差や意図せぬ解釈の揺らぎも許されないため、この厳密な性質は絶対条件である。

しかし近年、生産性向上の圧力に押された金融機関は、フロントエンドやミドルオフィスにMicrosoft 365 Copilotなどの生成AIアシスタントを急速に導入している。これらのLLM（大規模言語モデル）は、プロンプトの文脈や温度パラメータに応じて出力が変動する「確率論的（Probabilistic）」なシステムである。政府機関（例えば、AIセキュリティ要件を定めたS17のベースラインITセキュリティポリシーや、国際AI安全性レポート）がAIの限定的な利用と厳格な検証を推奨しているにもかかわらず、企業はこの非決定的なインターフェースを、決定論的なコアシステムと深く密結合させてしまった[27]。

Mythosのような高度な推論能力とハッキング能力を併せ持つAIが攻撃側に回った場合、この「確率と決定の境界線」が最大の標的となる。侵入のプロセスは以下のような連鎖を描く。第一段階として、攻撃側AIは従業員のCopilot環境に対し、巧妙に偽装された未知の不可視プロンプトインジェクションを実行し、RAGデータベースを論理的に汚染する。第二段階として、この汚染されたCopilotを通じて従業員の正規の認証トークンを密かに抽出し、権限の水平展開（Lateral Movement）を行って内部APIエンドポイントやSWIFTのメッセージ生成プロセスに到達する。第三段階において、Mythosは内部ネットワークで発見した未知の脆弱性（CVE-2026-4747に見られるような、認証不要のカーネルレベルRCE等）を悪用し、特権を昇格させてシステムの完全な制御権を奪取する[12]。最終段階として、完全に正規のフォーマットで偽装された大量の不正送金指示（SWIFTメッセージ等）を生成し、COBOLのコアシステムに流し込む。

ここで決定論的システムの脆弱性が露呈する。COBOLコアシステムやSWIFTは、入力された指示が「正規の認証チャネル」を経由していれば、その指示の文脈的な妥当性や経済的な異常性を検証する能力を持たない。結果として、機械的なスピードでグローバルな資金の移動が実行され、人間が異常に気づく頃には資金の汚染と流動性の枯渇が完了している。これは単なるデータ漏洩事件ではなく、資本市場の完全性に対する実存的な破壊を意味する。

## 3.2. 防御側の絶望的な非対称性とパッチサイクルの無効化

対象記事は、企業のパッチ適用期間の中央値（約70日）と、Mythosによる武器化の速度（数時間から数日）の間に存在する「時間的非対称性」に言及している[8]。この指摘は、過去数十年にわたって構築されてきたサイバーセキュリティの基本パラダイムが完全に崩壊したことを意味する。

これまで、サイバー防衛の基本方程式は「脆弱性が発見（公開）されてから、攻撃者がそれをリバースエンジニアリングして兵器化するまでの猶予期間（リードタイム）内に、パッチを適用するか一時的な緩和策を講じる」というものであった。Trend Microなどの主要なセキュリティ企業は、この猶予期間において最大96日間の早期保護（仮想パッチ）を提供することで、防衛線を維持してきた[18]。

しかし、MythosクラスのフロンティアAIの登場により、この時間軸は圧縮されたというよりも、事実上消滅した。AIは、公開された脆弱性を分析するだけでなく、「未知の脆弱性の発見」から「完全な自律的エクスプロイト（20以上のガジェットを連鎖させるようなROPチェーン等）の生成」までを、自律的かつ数分から数時間の単位で完了させる[16]。攻撃側がAIを活用することで、かつては高度なスキルを持つ人間の専門家チームが数ヶ月を要した攻撃手法の開発が、夜間に数十ドルのクラウドコンピューティングコストを費やすだけで自動生成されるようになったのである[15]。

この圧倒的な非対称性に対し、従来型のセキュリティ運用（SOC）や定期的なペネトレーションテストは全く追いつくことができない。防御側が70日かけてパッチを適用している間に、AIは数十の新たな侵入経路を発見し、自動的に攻撃を適応させてしまうからだ。この時間的非対称性への恐怖こそが、Anthropic社がMythosの一般公開を見送り、巨額のソフトウェア銘柄の暴落（SaaSpocalypse）を引き起こした根本的な原因である。

## 3.3. 単一栽培（Monoculture）アーキテクチャの脆弱性と増幅メカニズム

対象記事の第4項で論じられている「単一栽培（Monoculture）の危険性」は、システムの生態学的観点から極めて重要な洞察であり、金融危機がグローバルに波及するメカニズムを的確に説明している。

現在のエンタープライズITインフラは、フロントエンドにおけるMicrosoft製品（Windows, Active Directory, Microsoft 365 Copilot）への過度な依存によって特徴づけられている。同時に、バックエンドのサーバーやクラウドインフラストラクチャにおいては、Linuxカーネルや FreeBSDといった特定のオープンソースソフトウェアが世界的なモノカルチャーを形成している。

Mythosのような論理的推論能力に長けたAIにとって、モノカルチャー環境は「投資対効果が極めて高い」理想的な標的である。例えば、MythosがFreeBSDのNFSサーバーに17年間存在したCVE-2026-4747を発見しエクスプロイトを作成した際、それは単一のサーバーの脆弱性を見つけたにとどまらず、世界中で稼働する膨大な数のファイアウォール、ストレージアプライアンス、ネットワーク機器への侵入経路を同時に獲得したことを意味する[12]。

同様に、FFmpegのような普遍的なメディアコーデックライブラリの脆弱性や、クラウドのワークロードを分離する仮想マシンモニター（VMM）に存在するゲストOSからホストへのメモリ破損の脆弱性は、「一つの鍵で世界中のあらゆる金庫を開けられる」状態を作り出す[6]。対象記事が指摘するように、Microsoft Copilotの基盤に潜むRAGポイズニングの新たな脆弱性が一つでも発見されれば、それは世界の金融機関のフロントエンドを一斉に無力化し、システム全体に致命的な障害を連鎖させる

波及力を持つのである。

---

# 4. Project Glasswingとテクノロジー・金融業界の新たな防衛パラダイム

この絶望的な非対称性と構造的危機に対し、テクノロジー業界と金融業界はいかにして立ち向かおうとしているのか。対象記事でも言及されている「Project Glasswing」は、その解答の試みであり、前例のない防衛協定である。

## 4.1. アライアンスの目的とフロンティアAIによる先行者利益の確保

2026年4月7日にAnthropic社によって公に発表された「Project Glasswing」は、Amazon Web Services (AWS)、Apple、Broadcom、Cisco、CrowdStrike、Google、Microsoft、NVIDIA、Palo Alto Networks、Linux Foundation、そして金融界から唯一JPモルガン・チェースを初期パートナーとして迎え入れた共同イニシアチブである[6]。Anthropic社はこのプロジェクトに対し、最大1億ドルに上るMythosの利用クレジットと、オープンソースセキュリティ組織への400万ドルの寄付を提供している[29]。

このプロジェクトの真の戦略的意義は、**「悪意ある攻撃者が強力なAIを手にする前に、防御側のインフラプロバイダーにフロンティアAIによる自己診断とパッチ適用の先行者利益（ヘッドスタート）を与えること」**にある[3]。すなわち、AIによって圧縮された「攻撃側のリードタイム」に対抗するため、防御側もまたAIを用いて「コードベースの自律的な根本原因分析」と「脆弱性の発見・修正の高速化」を行うという、AI同士の軍拡競争のパラダイムである[32]。

## 4.2. 業界内の決定的な情報格差とシステミック・リスクの残存

しかし、この防衛アライアンスにも構造的な限界が存在する。前述の通り、財務省の緊急会合に出席した5つの巨大銀行（シティ、モルガン・スタンレー、バンク・オブ・アメリカ、ウェルズ・ファーゴ、ゴールドマン・サックス）は、Project Glasswingの初期リストに含まれておらず、現時点でMythosの能力を用いた自己システムの脆弱性診断を行うアクセス権を持っていない[2]。

これらの金融機関が管理する総額約9兆ドルの資産は、依然としてMythosクラスの能力によって発見されうる未知のゼロデイ攻撃に対して無防備な状態に置かれている。サイバー防御は「最も弱い環（チェーン）」から崩壊するという原則に従えば、JPモルガン・チェースのような一部の機関が極めて強固なAI防御盾を持っていたとしても、SWIFTのような相互接続されたグローバル金融ネットワーク全体の安全性を担保することはできない。この防御能力の非対称性こそが、財務省やFRBが抱く最大の懸念事項であり、緊急会合において全銀行に対する徹底的なペネトレーションテストやエンドポイントセキュリティの見直しが勧告された理由である[2]。

## 4.3. サイバー防衛ベンダーによる次世代AI防御の実装

この脅威に対し、サイバーセキュリティベンダーも座視しているわけではない。Trend Microなどの企業は、GenAIサービスに対する包括的な可視性と制御を提供する「AI Secure Access (ZTSA)」や、ブラインドスポットを排除するプロアクティブなサイバーセキュリティAI「Trend Cybertron」の導入を進

めている[18]。特に、CopilotやChatGPTのようなAIアシスタントに対するプロンプトと応答をリアルタイムで検査し、隠されたHTMLコメントや白文字によるプロンプトインジェクションを自然言語処理（NLP）で検知・ブロックする技術は、エコーリーク型の攻撃を防ぐための第一の防波堤となっている[18]。

さらに、NVIDIAやDell Technologiesとの協業による「Secure AI Factory」イニシアチブは、AIのライフサイクル全体を保護し、エンタープライズ環境におけるAIモデルの安全な基盤を構築することを目指している[18]。しかし、これらの防衛技術が、Mythosの自律的な学習能力とエクスプロイト生成速度を完全に凌駕できるかについては、依然として不確実性が高い。

---

# 5. 破局的シナリオを回避するための構造的変革と専門的推奨事項

対象記事の第5項では、金融システムの崩壊を防ぐための解決策として、構造的変化が提案されている。これらの提案は、最先端のシステムエンジニアリングおよびAI安全性の観点から見て、極めて妥当かつ不可避なアプローチである。本セクションでは、それらの提案をさらに深化させ、政策立案者と金融機関が直ちに実行すべき具体的な推奨事項を提示する。

## 5.1. マイクロセグメンテーションと疎結合（Loose Coupling）アーキテクチャへの完全移行

記事が提唱する「密結合から疎結合へ：システムを自己完結型の単位に分割し、波及を防ぐ」というアプローチは、現在の金融インフラ防衛における最優先課題である。

金融インフラにおける最大の弱点は、認証情報が一度突破されると、フロントエンドのAIアシスタントからバックエンドのSWIFTエンドポイントやCOBOLコアシステムまで、論理的にシームレスにアクセスできてしまうフラットなネットワーク設計にある。疎結合への完全移行とは、ゼロトラスト・アーキテクチャの原則を徹底し、各システム間にハードウェアレベル、あるいは厳格な論理レベルでの「マイクロセグメンテーション（細分化された防御壁）」と「APIゲートウェイ」を設けることを意味する。

トランザクションの要求とその実行プロセスを物理的・論理的に切り離すことで、仮にフロントエンドのCopilotが未知のゼロデイ攻撃によって完全に侵害されたとしても、被害はローカルな情報漏洩の範囲に封じ込められ、コアシステムへの不正な資金移動指示の注入に直結することを防ぐことができる。

## 5.2. エコシステムの多角化によるモノカルチャーの打破

「Microsoft依存からの脱却とオープンソース技術の組み合わせ」による多様性の確保という提案は、前述の「モノカルチャーのリスク」に対する直接的かつ根本的な処方箋である。

生物学的な生態系が遺伝的多様性を持つことで致死性のパンデミックに対する耐性を獲得するように、技術システムもベンダーやアーキテクチャの多様性を確保することで、AIによるゼロデイワームの世界的蔓延に対する「群免疫」を獲得することができる。特定のOS（Windows等）や特定のクラウドプロバイダーに全ての業務基盤を依存するのではなく、複数の異機種環境（マルチクラウド、ハイブリッドアーキテクチャ、異なるOSの採用）を意図的に混在させるシステム設計が求められる。これ

により、一つの決定的なゼロデイ脆弱性がシステム全体を連鎖的にダウンさせるリスクを数学的に低減することが可能になる。

### 5.3. AIエージェントの権限分離と動的シールドメカニズム（Human-in-the-Loop）の義務化

最も重要かつ即効性のある対策が、「AIの限定的利用：トランザクションの書き込み等には人間の最終承認を組み込む（プロダクトの自律的な核に据えない）」という設計思想の厳格な適用である。

AI研究と安全性の文脈において、自律型AIエージェントにデータを読み取る「Read」の権限を与えることと、データを変更・実行する「Write/Execute」の権限を与えることは、全く次元の異なるリスクを伴う。2025年に発表された国際AI安全性レポート（S17等で参照されるガイドライン）においても、機械学習システムのデータ駆動型で適応的な性質により、従来のテスト手法では包括的な安全性を保証できないことが強く指摘されている [28]。

金融機関は、AIアシスタントの役割をデータの検索、要約、分析といった「支援ツール（Read権限）」に厳格に制限すべきである。実際の資金移動、コアシステムへのデータ書き込み、権限の変更といった不可逆的なアクションについては、暗号論的な多要素認証（MFA）を用いた「人間の物理的な承認プロセス（Human-in-the-Loop）」をシステムレベルで強制しなければならない。

さらに、高度な対策として、S17の論文が提唱する「動的シールドメカニズム」の実装が不可欠となる。これは、AIエージェントの行動を監視・制限する別の「独立した監視用シールドAI」を配置し、エージェントの行動や協調状態に応じてシールドが動的に分割・結合することで、システムの安全性を自律的に維持するアプローチである [28]。攻撃側のAIが高度化する以上、防御側の監視機構にも同等以上の推論能力を持つAIを配置し、相互監視による安全性の担保を図る必要がある。

---

## 6. 結論：不可逆的な時代の転換点と金融レジリエンスの再定義

本報告書による厳格な検証と深層分析の結果、対象記事（mythos-copilot-financial-crisis-final.md）は、エコーリーク（CVE-2025-32711）の時系列的な適用性に関する一部の技術的齟齬を含んでいるものの、その中心的な主張と提示された崩壊シナリオの蓋然性は極めて高く、むしろ現実の脅威の深刻さを極めて正確に捉えていると評価される。

2026年4月にAnthropic社が直面し、米国財務省およびFRBが異例の緊急介入を余儀なくされた「Claude Mythos Preview」の脅威は、サイバーセキュリティの歴史において「人間の認知能力と防御スピードによってAIの脅威を統制できた時代」の完全な終焉を意味する。AIモデルが、17年間人間の目に見つからなかったカーネルレベルの脆弱性を自律的に特定し、数時間で完全なエクスプロイトチェーンを構築できるようになった現在、防御側に残されていた「パッチ適用のリードタイム」は消滅した。

世界の金融業界は、COBOLという半世紀前のレガシーシステムがもたらす技術的負債と、AIアシスタントの無秩序な統合という性急な近代化の間に挟まれ、歴史上最も脆弱な過渡期に直面している。JPモルガン・チェースなど一部の機関がProject Glasswingを通じて辛うじて防衛の最前線に

立っている一方で、業界全体としての対応は致命的に遅れており、SaaSpocalypseが示した市場の恐怖は極めて合理的な反応である。

対象記事が警告するように、真のシステミック・リスクの顕在化を防ぐためには、AIの利便性に盲目的に依存する現在のアプローチを即座に放棄しなければならない。「疎結合化によるシステムの分断」「エコシステムの多様性確保」、そして「厳格な人間の介入プロセス（Human-in-the-Loop）の再構築」という、短期的には効率性を犠牲にする痛みを伴う根本的なアーキテクチャの再設計へと舵を切る必要がある。Mythosが世界に突きつけたのは、単なる新しいハッキングツールの登場ではなく、グローバルインフラストラクチャに対する「人間の決定論的統治」の喪失という、実存的な危機に他ならない。金融システムのレジリエンスは今、機械の推論速度との容赦ない競争の中で、その根本からの再定義を迫られている。

## 引用文献


1. Anthropic's Claude Mythos AI fears trigger $2 trillion wipeout in IT ..., 4月 12, 2026にアクセス、https://m.economictimes.com/news/new-updates/anthropics-claude-mythos-ai-fears-triggers-2-trillion-wipeout-in-it-stocks-jpmorgan-ceo-jamie-dimon-warns-ai-will-likely-/articleshow/130187154.cms
2. Anthropic's AI, the most powerful on Earth, prompted an emergency meeting on Wall Street, but JPMorgan Chase, which had the "cure," was absent. | 律动 BlockBeats on Binance Square, 4月 12, 2026にアクセス、https://www.binance.com/en/square/post/310884313815841
3. Bessent and Powell send Wall Street's biggest banks a warning - TheStreet, 4月 12, 2026にアクセス、https://www.thestreet.com/economy/bessent-and-powell-send-wall-streets-biggest-banks-a-warning
4. Treasury Proposes Rule to Implement the GENIUS Act’s Requirements to Counter Illicit Finance, 4月 12, 2026にアクセス、https://home.treasury.gov/news/press-releases/sb0435
5. FinCEN Proposes Rule to Fundamentally Reform Financial Institution Programs Designed to Fight Illicit Finance, 4月 12, 2026にアクセス、https://www.fincen.gov/news/news-releases/fincen-proposes-rule-fundamentally-reform-financial-institution-programs
6. Anthropic Launches Project Glasswing to Use AI to Find and Fix Critical Software Vulnerabilities, 4月 12, 2026にアクセス、https://www.infosecurity-magazine.com/news/anthropic-launch-project-glasswing/
7. An AI Model Scared the Federal Reserve and US Treasury Into Summoning Wall Street, 4月 12, 2026にアクセス、https://aimmediahouse.com/ai-bfsi/an-ai-model-scared-the-federal-reserve-and-us-treasury-into-summoning-wall-street
8. mythos-copilot-financial-crisis-final.md
9. Overnight, will countless bugs appear on your phone and computer? - 36氪, 4月 12, 2026にアクセス、https://eu.36kr.com/en/p/3758841606079234

10. Claude Mythos: Benchmark-Dominating AI with Real Risks - Labellerr, 4月 12, 2026にアクセス、https://www.labellerr.com/blog/anthropic-claude-mythos-capabilities/
11. Claude Mythos: The AI Model Anthropic Built (and Refused to Release) - Thesys, 4月 12, 2026にアクセス、https://www.thesys.dev/blogs/claude-mythos
12. CVE-2026-4747: FreeBSD RPCSEC_GSS RCE Vulnerability - SentinelOne, 4月 12, 2026にアクセス、https://www.sentinelone.com/vulnerability-database/cve-2026-4747/
13. CVE-2026-4747 | Mondoo Vulnerability Intelligence, 4月 12, 2026にアクセス、https://mondoo.com/vulnerability-intelligence/vulnerability/CVE-2026-4747
14. Anthropic's new AI model finds and exploits zero-days across every major OS and browser, 4月 12, 2026にアクセス、https://www.helpnetsecurity.com/2026/04/08/anthropic-claude-mythos-preview-identify-vulnerabilities/
15. Mythos autonomously exploited vulnerabilities that survived 27 years of human review. Security teams need a new detection playbook, 4月 12, 2026にアクセス、https://venturebeat.com/security/mythos-detection-ceiling-security-teams-new-playbook
16. Claude Mythos Preview \ red.anthropic.com, 4月 12, 2026にアクセス、https://red.anthropic.com/2026/mythos-preview/
17. Anthropic announces 'Project Glasswing' in alliance with tech giants to strengthen global cybersecurity, 4月 12, 2026にアクセス、https://www.aninews.in/news/business/anthropic-announces-project-glasswing-in-alliance-with-tech-giants-to-strengthen-global-cybersecurity20260408083637
18. Preventing Zero-Click AI Threats: Insights from EchoLeak | Trend Micro, 4月 12, 2026にアクセス、https://www.trendmicro.com/en/research/25/g/preventing-zero-click-ai-threats-insights-from-echoleak.html
19. How Microsoft 365 Copilot Customers Should Think About Agent Control - Rubrik, 4月 12, 2026にアクセス、https://www.rubrik.com/blog/company/26/4/how-microsoft-365-copilot-customers-should-think-about-agent-control
20. 'EchoLeak' AI Attack Enabled Theft of Sensitive Data via Microsoft 365 Copilot, 4月 12, 2026にアクセス、https://www.securityweek.com/echoleak-ai-attack-enabled-theft-of-sensitive-data-via-microsoft-365-copilot/
21. EchoLeak in Microsoft Copilot: What it Means for AI Security - Varonis, 4月 12, 2026にアクセス、https://www.varonis.com/blog/echoleak
22. CVE-2025-32711 Detail - NVD, 4月 12, 2026にアクセス、https://nvd.nist.gov/vuln/detail/cve-2025-32711
23. DXC's Mainframe Engineering Services team helps banks master COBOL core banking, 4月 12, 2026にアクセス、https://dxc.com/insights/knowledge-base/blogs/why-banks-still-rely-on-cobol-driven-mainframe-systems

24. 44 Legacy System Modernization Statistics Every Enterprise Should Know in 2026, 4月 12, 2026にアクセス、https://www.dreamfactory.com/hub/legacy-system-modernization-statistics
25. Learn how progressive modernization enables banks to transform core systems without big bang risk. - Softjourn, 4月 12, 2026にアクセス、https://softjourn.com/insights/core-banking-modernization-in-5-steps
26. Legacy Banking Costs: The Hidden Financial and Strategic Burden on Financial Institutions - FinTech Consultant, 4月 12, 2026にアクセス、https://www.fintech-consultant.com/post/hidden-costs-of-legacy-banking-infrastructure-on-financial-institutions
27. Information and Cyber Security Within the Government | Digital Policy Office, 4月 12, 2026にアクセス、https://www.digitalpolicy.gov.hk/en/our_work/digital_infrastructure/information_cyber_security/government/
28. Formal methods for safety-critical machine learning: a systematic literature review - Frontiers, 4月 12, 2026にアクセス、https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2026.1749956/full
29. Project Glasswing - Anthropic, 4月 12, 2026にアクセス、https://www.anthropic.com/project/glasswing
30. Claude Mythos, Anthropic AI capable of hacking any software, joins forces with Google, Apple, AWS & more; Users' personal data at risk?, 4月 12, 2026にアクセス、https://m.economictimes.com/news/new-updates/claude-mythos-anthropic-ai-capable-of-hacking-any-software-joins-forces-with-google-apple-aws-more-users-personal-data-at-risk/articleshow/130106401.cms
31. Anthropic launches Project Glasswing, says Claude Mythos found risks in every major OS and browser, 4月 12, 2026にアクセス、https://www.financialexpress.com/life/technology-anthropic-defines-project-glasswing-says-mythos-has-found-vulnerabilities-in-thousands-of-systems-4201201/
32. Why Project Glasswing Marks a Turning Point for Cybersecurity | Arctic Wolf, 4月 12, 2026にアクセス、https://arcticwolf.com/resources/blog/project-glasswing-marks-a-turning-point-for-cybersecurity/
33. Policymaker's Guide to International AI Safety Coordination | Digital Watch Observatory, 4月 12, 2026にアクセス、https://dig.watch/event/india-ai-impact-summit-2026/policymakers-guide-to-international-ai-safety-coordination/