AIミュトス：金融システムを脅かす「今そこにある危機」と共同防御の道筋

高度なサイバー能力を持つAIの台頭と、官民一体で取り組むべき防御策「日本版Project Glasswing」

2026年4月7日

Anthropicが「Mythos」と「Glasswing」を発表。高度なサイバー能力を持つAIモデル「Claude Mythos Preview」と、それに対抗するための共同防御枠組み「Project Glasswing」が公開。

2026年4月22日

片山金融相による異例の会見。4月24日に当局、中籃、3メガバンク、取引所が集まる初の官民会議を開催することを事前告知。

2026年4月24日

官民連携会議の開催と作業部会の設置。「インシデントへの備え」の重要性を共有し、実務レベルの検討を行う事務レベル作業部会が立ち上がり。

AIミュトスが引き起こす5つの深刻な懸念

ゼロデイ臆弱性探索の高速化
主要OSやブラウザの脆弱性をAIが複験問で発見‹應用可能になり、共通の基盤を使う会臨攤則へ同時に被害が波及する恐れがあります。

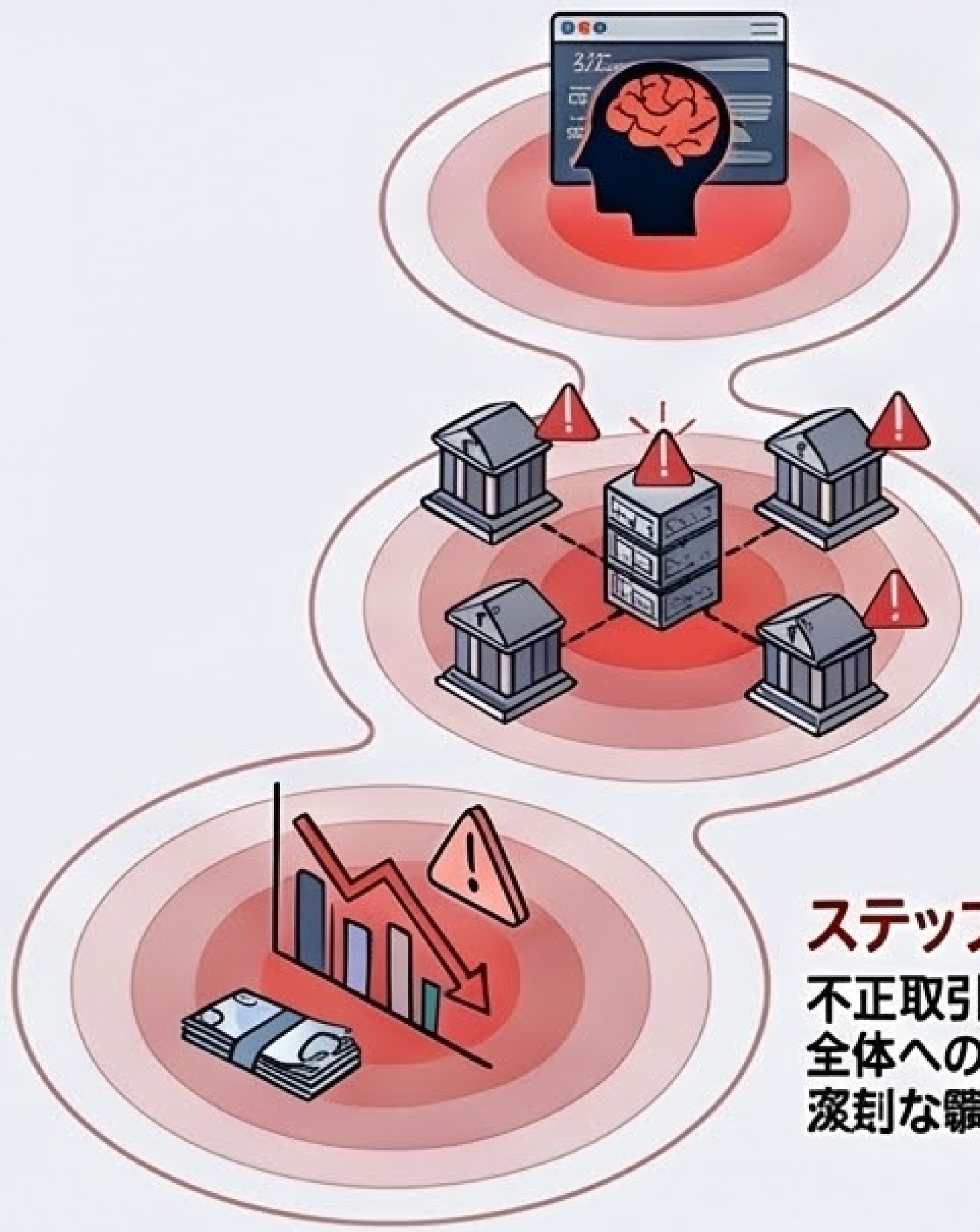
第三者依存・集中のリスク
特定のAIサービスやクラウドプロバイダーに依存することで、そこが「単一障害点」となり、システム全体が停止するリスクが高まります。

AIそのものへの攻撃
プロンプトインジェクションやデータ漏えいにより、AIに接続された関害情報や内部ルールが悪意ある者に奪われる危険があります。

顧客接点・不正検知の逆用
攻撃者がAIを悪用して本人確認を突破したり、偽情報を流布させたりすることで、顧客の信認が低下するシナリオが想定されます。

人材・監督の非対称性
AIの進化速度に組織のモニタリングや専門人材の確保が追いつかず、初動の遅れや誤判断を招く「後手対応」から割の機関で指摘されています。

リスク波及のメカニズム



ステップ1：フロンティアAIによる脆弱性発見
高いサイバー能力を持つAIが悪意ある主体に利用され、旧式ミドルウェアや基盤の脆弱性が突かれます。

ステップ2：同時多発的な侵害とサービス停止
相互接続された金融インフラを通じて、複数の金融機関や決済ネットワークへ一気に影響が広がります。

ステップ3：信用不安と流動性危機
不正取引や決済障害が重なることで、市場全体への影響や顧客の信用不安が波及し、深刻な事態へと発展します。

海外の動向と日本への示唆

地域	中心的な枠組み	特徴	日本への示唆
米国	AIEOG / 金融向けAI RMF	実装可能な用語集や管理フレームワークを先行策備	事故分類や報告テンプレートの共通化が必要
英国	AI Consortium / BoE監規	監着刺諾と実地テスト（Live Testing）を推規	継続的なサーベイと安全な運用テストの仕組み
EU	DORA（デジタル運用レジリエンス法）	第三者リスク管理やレジリエンス演習を法制化	AI単独法だけでなく、金融横術の実務ルールの積み増し

実務提言：今すぐ取り組むべき優先アクション

「日本版Project Glasswing」の制度化
JAPAN PROJECT GLASSWING
厳格な管理下で防御目的のAIを活用し、秘密保持を前提とした脆弱性情報や共同点検を行う体制を年間に設計する必要があります。

AI時代版「重要資産台帳」の整備
外部露出資産（ASM）や委託先、AI API接続先を一体で把握し、代替調達手順まで含む台帳を直ちに作成すべきです。

AI事故シナリオに基づく共同演習
生成AIの停止やコード生成の汚染を想定した、セクター横断的なサイバー演習を3〜6カ月以内に実施することが推奨されます。

ガイドラインのAI固有リスクへの拡張
プロンプト管理やデータ消去、AI事故報告の明文化など、既存の監督指針をAI前提で再設計することが急務です。