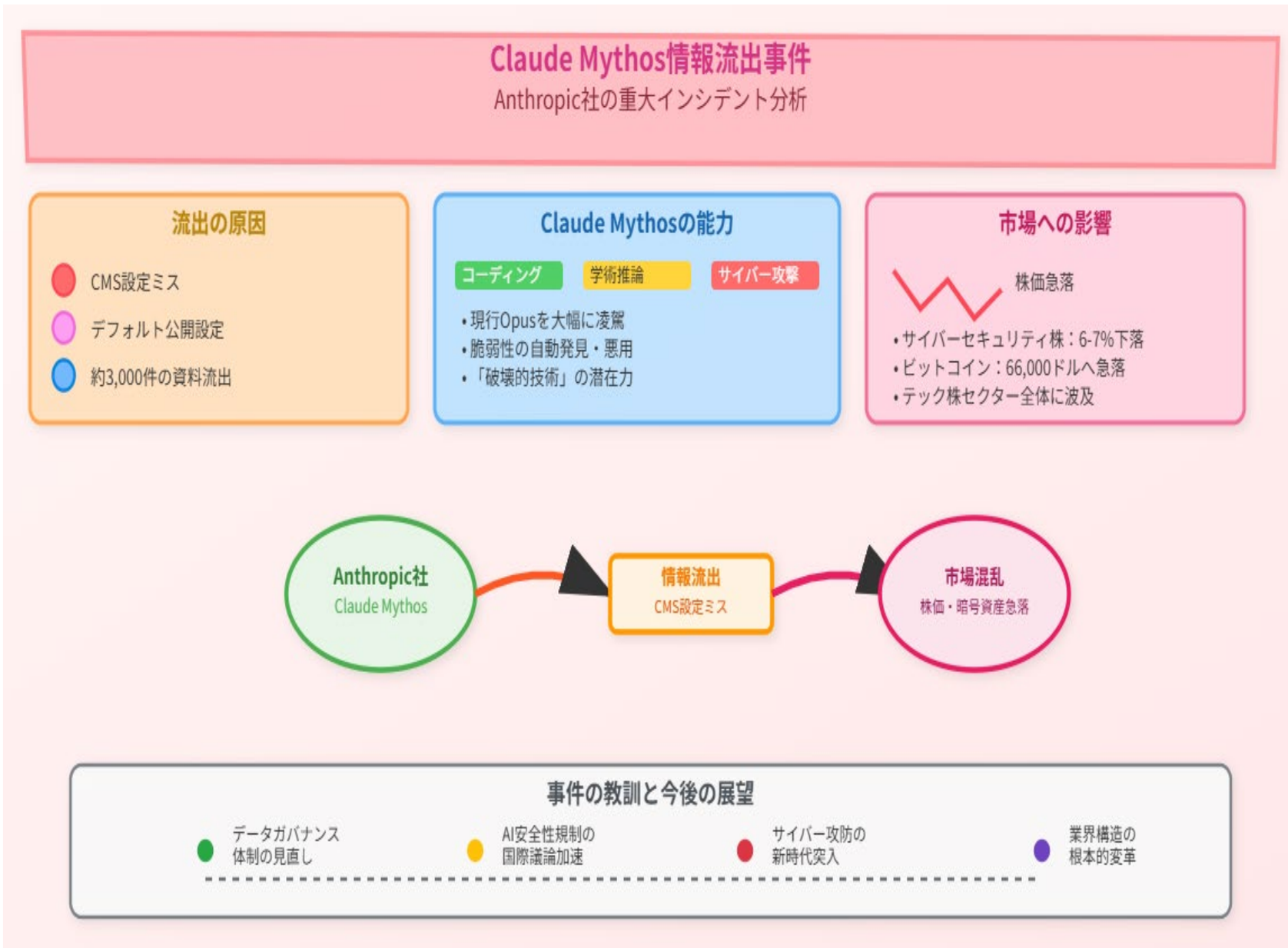
「Anthropic から新 AI「Claude Mythos」関連文書が流出」問題

Felo AI



概要

AI 開発企業 Anthropic 社において、コンテンツ管理システム（CMS）の設定不備という人為的ミスにより、未発表の次世代 AI モデル「Claude Mythos」に関する内部文書が外部からアクセス可能な状態となる重大な情報流出インシデントが発生しました 5 17 。この流出により、現行の最上位モデル「Opus」を大幅に凌駕する Mythos の驚異的な性能、特にソフトウェアの脆弱性を自動で発見・悪用しうる卓越したサイバーセキュリティ能力が白日の下に晒されました 1 3 9 。この事実は、AI によるサイバー攻撃の脅威を現実のものとして市場に強く意識させ、サイバーセキュリティ関連銘柄や暗号資産市場の急落を引き起こすなど、広範囲にわたる経済的影響を及ぼしました 1 6 25 。この事件は、AI 開発の最前線を走る企業のデータガバナンスの脆弱性と、急速に進化する AI がもたらす「破壊的技術」としてのリスクを浮き彫りにした歴史的な出来事です 5 6 。

詳細レポート

流出の経緯と原因：初歩的な人為的ミス

今回の情報流出は、外部からのハッキングによるものではなく、Anthropic 社内の初歩的な設定ミスに起因します 17 。Fortune 誌の報道によって発覚したこの問題の原因は、同社が使用する CMS が、アップロードされたデジタルアセットを「デフォルトで公開状態」に設定していたことにありました 5 8 。この基本的な設定が見過ごされた結果、ユーザーが明示的に非公開設定を行わない限り、ファイルに公開 URL が自動で割り当てられ、約 3,000 件に及ぶ内部資産（ブログ投稿の草稿、画像、PDF ファイルなど）が誰でも閲覧可能な状態でインターネット上に放置されていました 1 5 17 。セキュリティ研究者の指摘を受け、Anthropic は直ちに問題のアセットへのアクセスを制限しました 17 。同社は、流出したのは初期の草稿段階の資料であり、顧客データや AI のコアシステムへの不正アクセスは確認されていないと説明しています 17 。「安全性第一」を企業理念の中核に据える Anthropic が、高度な AI ではなく基本的な人為的ミスによって自社の最重要機密を漏洩させたという事実は、AI 開発企業におけるデータガバナンスの重要性を改めて突きつける結果となりました 5 8 。

未発表モデル「Claude Mythos」の驚異的な能力

流出した文書により、これまで未公開だった次世代 AI モデル「Claude Mythos」の存在と、その詳細な能力が明らかになりました [1] [4] [5]。このモデルは、社内で「Capybara」と呼ばれる新たな最上位ティアに属しており、現行の最上位モデルである Claude Opus を大幅に上回る性能を持つとされています [5] [6] [17]。Anthropic 自身の内部文書では、Mythos を「パフォーマンスにおける画期的な変化」「これまでに構築した中で最も高性能なもの」と評価しています [1]。
具体的には、Claude Opus 4.6 と比較して、以下の分野で劇的に高いスコアを記録していることが示されています [2] [7] [9]。

- **ソフトウェアコーディング**
- **学術的推論**
- **サイバーセキュリティ**

特にサイバーセキュリティ分野における能力は突出しており、内部文書には「他のあらゆる AI のサイバー能力を凌駕する」と明記されていました [3] [9]。このモデルは、現在一部のサイバー防衛組織などを対象に、その挙動とリスクを評価するための早期アクセステストが行われています [1] [17]。

市場への衝撃とサイバーセキュリティのパラダイムシフト

「Claude Mythos」が持つサイバーセキュリティへの潜在的脅威が市場に与えた衝撃は甚大でした。流出文書には、Mythos が「ソフトウェアの脆弱性を迅速に発見し、悪用する能力を持つ」可能性が警告されており、「防御者の努力をはるかに上回るペースで脆弱性を悪用できるモデルの波を引き起こす恐れがある」との深刻な懸念まで記されていました[1][9]。さらに、このモデルが潜在的な「破壊的技術（destructive technology）」となりうることが示唆されており、サイバー兵器開発競争を加速させる危険性が指摘されています[1][6]。

この報道を受け、金融市場は即座に反応しました。

- **株式市場**: AI によるサイバー攻撃の脅威が現実のものとして認識され、Palo Alto Networks が 6〜7%下落したほか、CrowdStrike などの主要なサイバーセキュリティ関連株が軒並み値を下げました[13][17]。テクノロジー・ソフトウェアセクター全体を示す iShares Expanded Tech-Software Sector ETF (IGV) も約 3%下落しました[1]。
- **暗号資産市場**: 市場全体の不安心理が波及し、ビットコイン価格が 66,000 ドルへ急落する一因となった可能性が指摘されています[1][16]。

この事件は、AI とサイバーセキュリティの「軍拡競争」が新たな段階に突入したことを示しています[9]。短期的には、防御側も Mythos クラスの AI を活用した次世代のセキュリティ製品開発を加速させる必要に迫られます[9]。中長期的には、「AI が脆弱性を発見し、AI がそれを防ぐ」という新たな攻防の時代が本格化し、この変化に迅速に適応できる企業が次世代の覇権を握ると考えられます[9]。

事件が示す教訓と今後の展望

Claude Mythos の流出事件は、単なる一企業のセキュリティインシデントにとどまりません。AI の能力が既存産業の構造を根底から覆しうることをリアルタイムで証明した歴史的な出来事です[9]。意図せずして「AI の核兵器の設計図」を世界に公開してしまったともいえるこの事例は、AI 開発企業に対し、より厳格な情報管理体制と公開ポリシーを求める声を強める契機となるでしょう[9]。

Anthropic をはじめとする AI 企業は、自らが開発する技術の潜在的なリスクを深く理解し、その管理体制を徹底的に見直すことが急務です。また、この事件は AI の安全性に関する国際的な議論に新たな問題提起をもたらし、高能力 AI モデルの取り扱いに関する規制やガイドライン策定の動きを加速させる可能性があります[9]。今後、AI 開発企業の情報管理の在り方と、サイバーセキュリティ業界の AI 適応戦略が、これまで以上に厳しく問われる時代が到来します[9]。

1. Anthropic の大規模な『Claude Mythos』リークにより
2. Anthropic「Claude Mythos」流出で AI 懸念が高まりサイバー ...
3. 未発表モデル「Claude Mythos」が「他のあらゆる AI のサイバー ...

4. なるほど、Opus（作品）の上だから Mythos（神話）なんだね。
5. Anthropic、未発表の最強 AI「Claude Mythos（Capybara） ...
6. アンソロピックの破壊的 AI「Claude Mythos」資料流出
7. 【速報】Anthropic の次世代モデル 「Claude Mythos」の存在が ...
8. Claude が「3 月だけ」でやったこと多すぎる。Mythos・Auto ...
9. Anthropic が AI の「核兵器」を誤公開──Claude Mythos と ...
10. Anthropic の新しいロボットモデルがサイバーセキュリティの懸念 ...
11. Claude Code にバックドア入り OSS を渡したら、何の疑いもなく ...
12. 未発表の最強 AI「Claude Mythos（Capybara）」の存在 ...
13. Anthropic の Claude Mythos モデル公開でサイバーセキュリティ ...
14. 伊藤穣一氏が声明発表 エプスタイン氏巡る報道「事実誤認 ...
15. 「Claude Cowork」にファイル流出の脆弱性 巧妙な間接的 ... － IT
16. アンソロピック「クロード神話」リークが仮想通貨市場を揺るがす
17. アンソロピックの未公開 AI 流出｜株式や仮想通貨市場に警戒感
18. アンソロピックの破壊的 AI「Claude Mythos」資料流出
19. レイモンド・ジェームズ、強固なファンダメンタルズにもかかわら ...
20. Anthropic が最強 AI「Claude Mythos」を誤漏洩 － 情報の灯台
21. 神話が始まる。リークされた次期 AI「Claude Mythos」がヤバ ...
22. 仮想通貨収益の非開示疑惑めぐる集団訴訟が正式認定
23. この反逆的なトランプの陰謀があなたの家庭に浸透している
24. Claude 悪用でメキシコ政府から 150GB 流出、1 億 9500 万件の ...
25. CoinPost｜仮想通貨ビットコインニュース・投資情報
26. Claude 利用率が 1487%急増、AI 競争が激化
27. 「深刻な懸念」を伴うサイバーセキュリティリスクが、Anthropic の ...