次世代AI「Claude Mythos」情報流出事件：人為的ミスが露呈させた破壊的ポテンシャル

Anthropic社のCMS設定不備により、未発表の次世代AI「Claude Mythos」に関する内部文書約3,000件が流出。
現行モデルを凌駕する高いサイバー攻撃能力が判明し、市場に大きな動揺を与えた。

インシデントの原因と概要

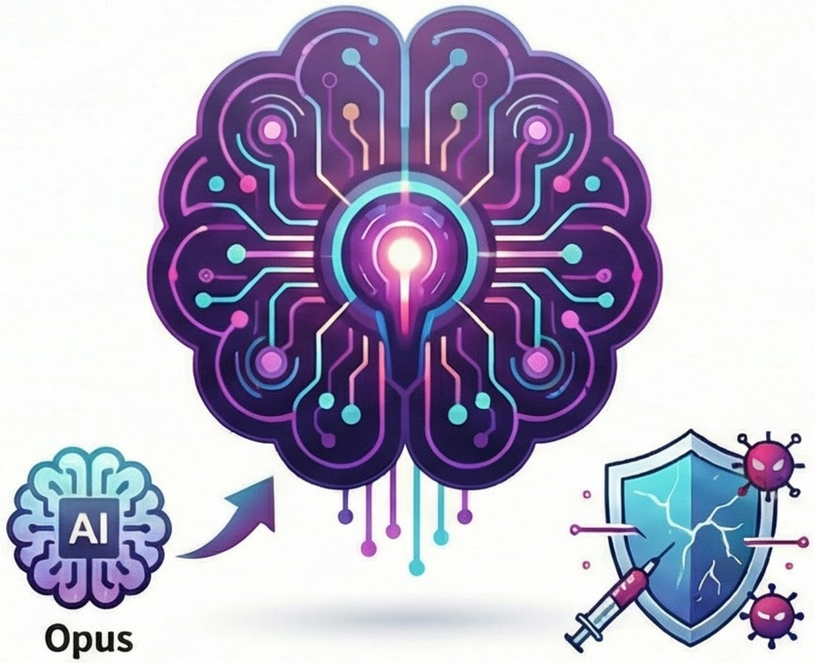
初歩的なCMS設定ミス

デジタルアセットがデフォルトで「公開状態」に設定される人為的ミスが原因。

約3,000件の内部資料が流出

ブログ草稿、画像、PDFなど、機密性の高い内部資産が一時閲覧可能となった。

次世代モデル「Claude Mythos」の正体



現行最上位「Opus」を大幅に凌駕

ソフトウェア開発、学術的推論、サイバーセキュリティの全分野で劇的に高い性能を記録。

脆弱性を自動発見・悪用する能力

既存のAIを凌駕するサイバー能力を持ち、「破壊的技術」としての潜在性が判明。

市場への衝撃と今後の教訓

金融市場・暗号資産の急落

サイバー攻撃の脅威が意識され、セキュリティ関連株やビットコインが大幅に値を下げた。

AIガバナンスの再定義が急務

「AIの核兵器」とも言える技術の管理には、より厳格な情報管理体制と規制が必要。