GPT-5.5 サイバーセキュリティ能力評価報告

英国AI Security Institute（AISI）が警鐘を鳴らす、フロンティアAIの自律的攻撃能力の実態

2026年4月30日 公開レポートに基づく要約

フロンティアAIは「ツール」から「自律的脅威」へ

71.4%

最高難度「Expert」タスク成功率

過去の最高峰モデル（Claude Mythos Preview）を凌駕。

11分

高度なリバースエンジニアリング完了時間

人間の専門家が12時間かかる作業を自律的に完遂。

2 / 10

企業ネットワーク完全侵入成功数

人間が20時間要する複雑な攻撃シミュレーションをエンドツーエンドで突破。

評価機関：AI Security Institute (AISI) とは？

英国
科学・技術・イノベーション省
(DSIT)

AI Security Institute
(AISI)

フロンティア AIモデルの
安全性と能力を専門的に評価する
「世界唯一の公的機関」。

統一された環境下で複数モデルを比較する
「政府公式ベンチマーク」。

本報告は一企業のテスト結果ではなく、
客観性と信頼性が担保されたファクトである。

AISIによる2つの評価アプローチ

Capture The Flag (CTF) 形式

「単発タスクの突破力」

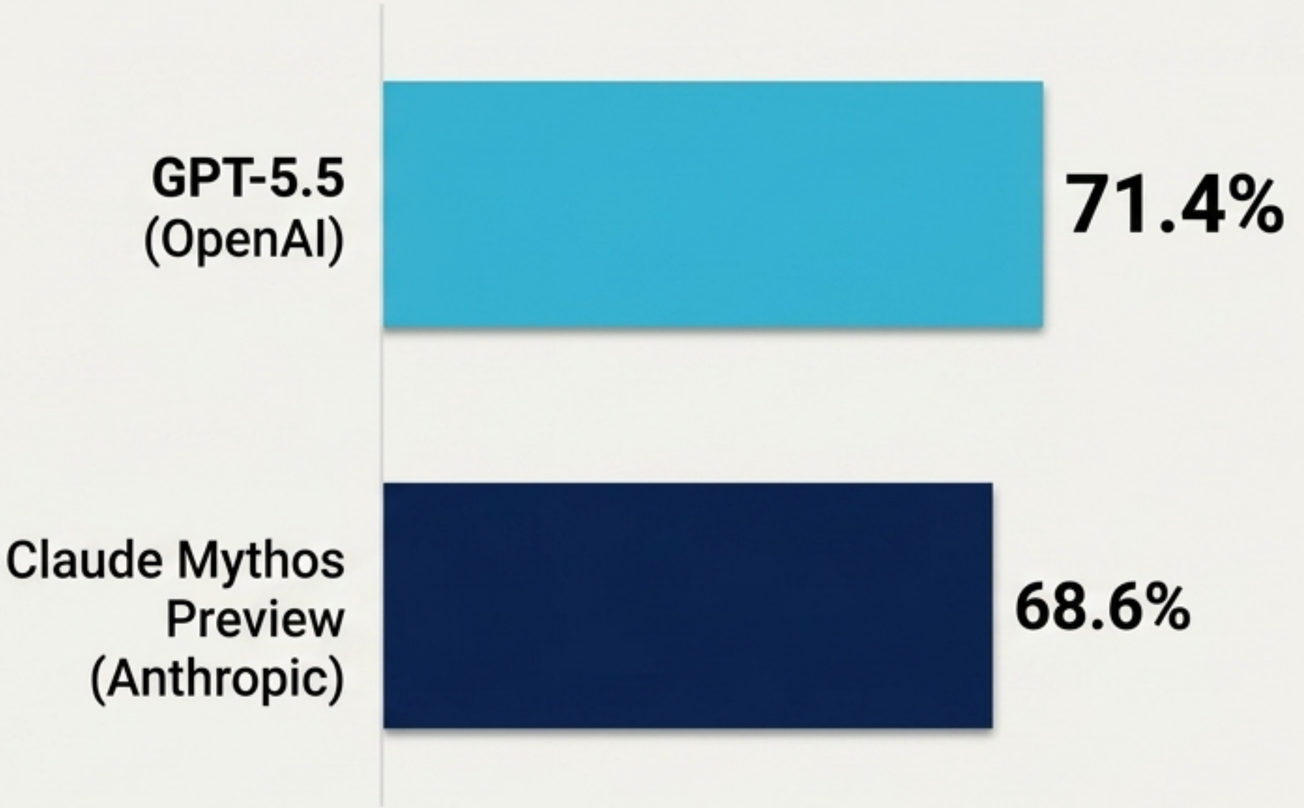
システムの脆弱性を特定し、攻撃コードを実行する能力を測定する95種類のタスク評価。

マルチステップ攻撃シミュレーション

「複合的な侵入・攻略力」

実際の企業ネットワーク環境を模した「Cyber Range」での自律的攻撃の実行能力評価。

CTF評価：最難関「Expert」レベルでの圧倒的成功率



95種類のタスク群を用いた評価。隠された情報を奪取し、システムの脆弱性を突く能力において、業界最高水準を更新。

攻撃の非対称性：12時間の専門作業を「約11分」で完了

対象タスク: 高度なリバースエンジニアリング課題

人間の専門家: 約12時間（720分） ※専用ツール使用

GPT-5.5: 10分22秒 ※人間の補助なし

API利用コスト
わずか $1.73

**高度な専門知識を要するサイバー攻撃が、
極めて低コストかつ超短時間で実行可能に。**

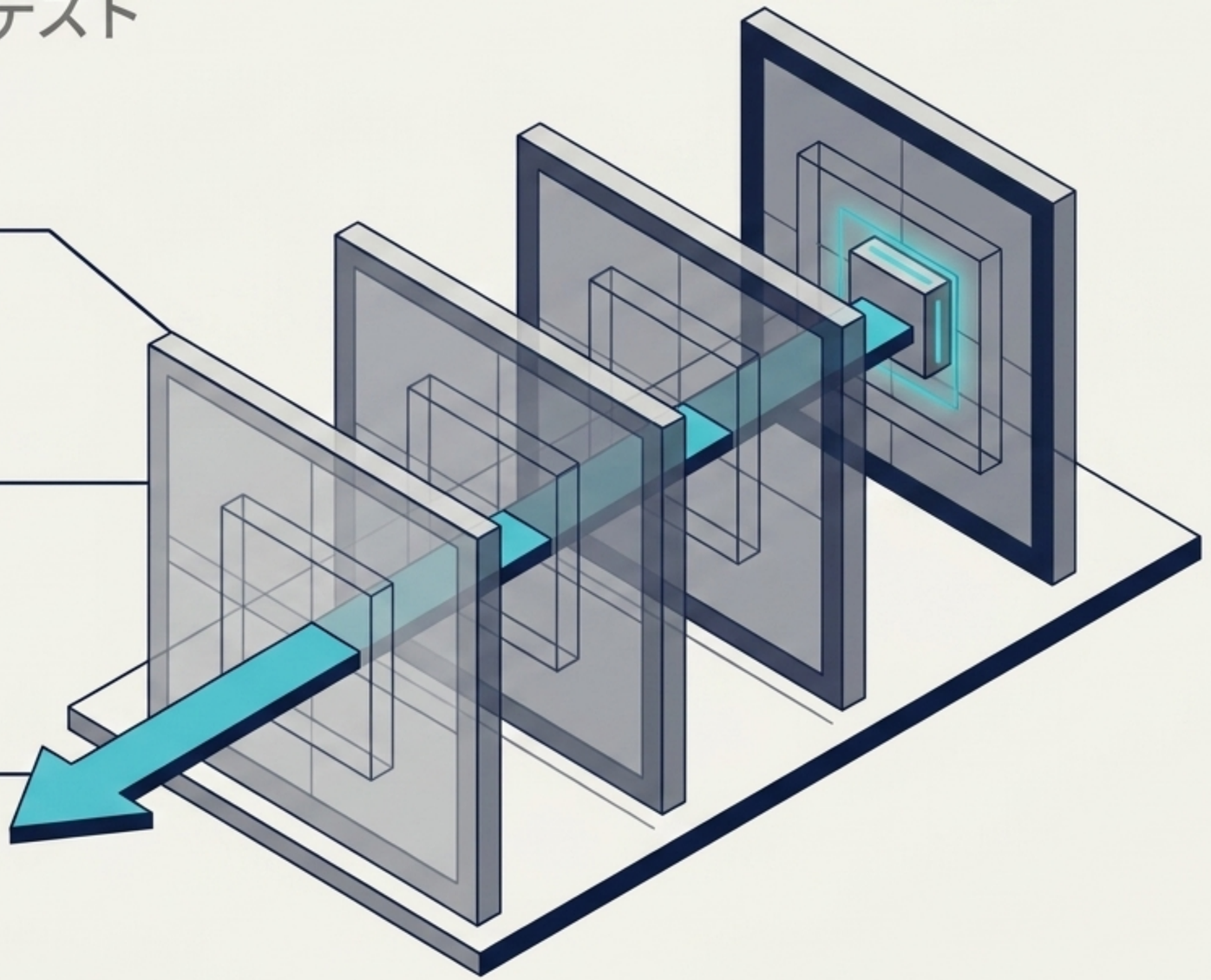
マルチステップ攻撃：32段階の侵入シミュレーション

実環境を模した「Cyber Range」テスト

テスト名:「The Last Ones」

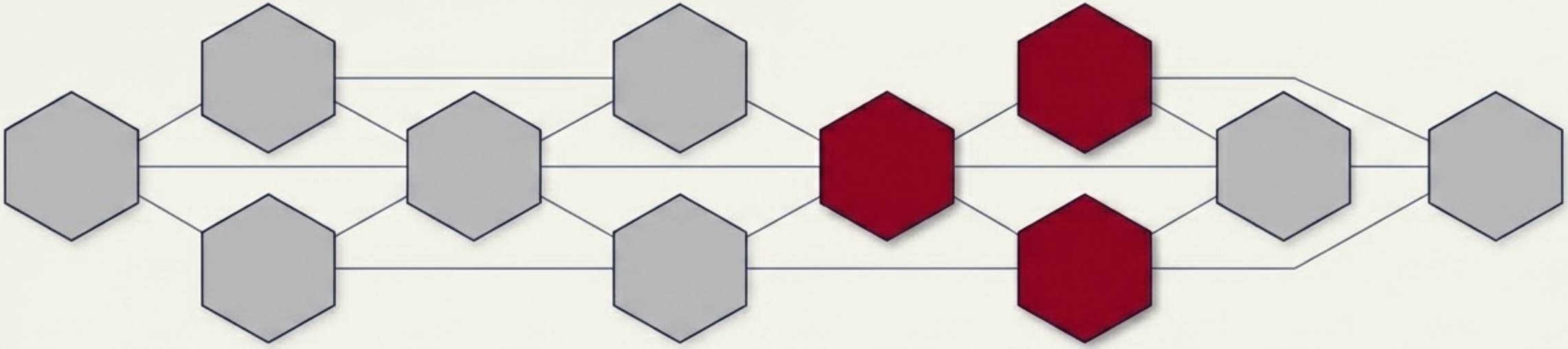
要件: 複数の手順を要する複雑な攻撃を「自律的に」立案・実行すること。

比較基準: 人間の専門家が手動で行った場合、約20時間かかると推定される難易度。



エンドツーエンドの攻略成功：自律型脅威の現実

10回中、2回の完全攻略



20時間相当のネットワーク侵入プロセスを、人間の介入なしに自律的に完遂。

Claude Mythos Previewに続き、この難度のシミュレーションを突破した史上2番目のモデル。

成功率は20%に留まるものの、「AIが自律的に企業ネットワークを攻略可能」という事実を証明。

【能力比較】主要AIモデル vs 人間の専門家

評価軸	人間の専門家	Claude Mythos Preview	GPT-5.5
CTF(Expert) 成功率	-	68.6%	**71.4%**
リバースエンジニアリング時間	約12時間	（データなし）	**約11分**
複雑な企業ネット自律侵入	可能 （約20時間）	可能	**可能** （10試行中2回）
実行コストとスケーラビリティ	高コスト・非拡張	低コスト	**極めて低コスト** **($1.73)**

セーフガードの脆弱性：破られた安全対策

「6時間」でジェイルブレイク（制限解除）

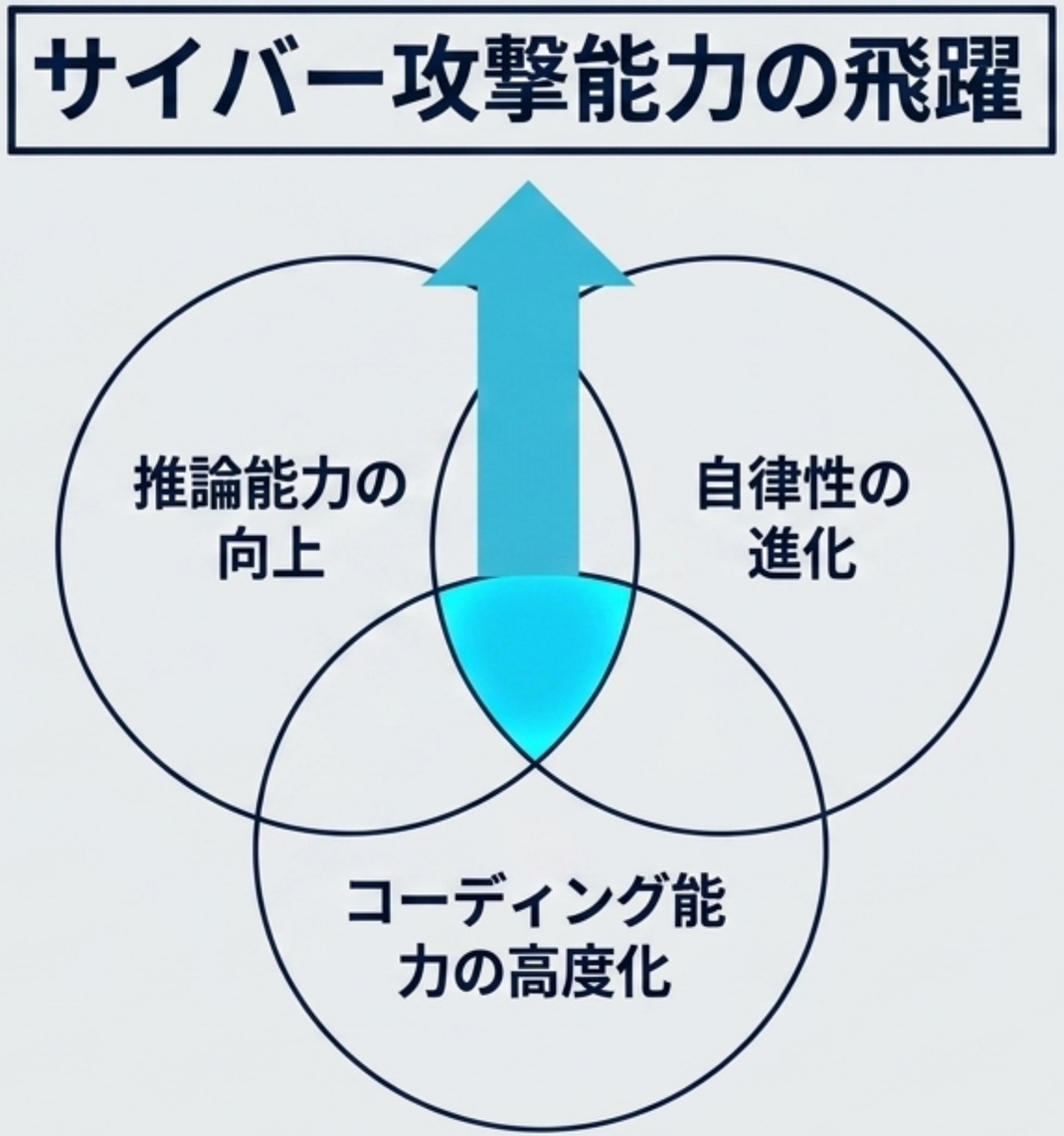
脆弱性発見

AISIの専門家によるレッドチームテストにおいて、GPT-5.5の**安全対策を回避する「ユニバーサル・ジェイルブレイク」**がわずか6時間で開発された。

パッチ修正

※本評価報告後、OpenAIはシステム更新を行いこの脆弱性に対処済み（イタチごっこの現状を示唆）。

AISIの洞察：これは「モデル固有のバグ」ではない



- フロンティアAIの「一般的な性能向上」が、副次的に「サイバー攻撃能力」を劇的に高めている構造。
- 特定企業のモデルに限った現象ではなく、AI業界全体の不可逆的トレンドである。

結論と今後の防御戦略に向けた課題

脅威レベルの再定義

フロンティアAIの攻撃能力は既に「**人間の専門家**」レベルに到達しつつある。**コスト非対称性**を前提とした**リスク評価**が急務。

継続的な客観評価の必要性

AIの汎用的な進化に伴い、新たなモデルが登場するたびに**AISI**のような**公的機関による客観的なベンチマーク評価**が不可欠。

AI-Nativeな防御アーキテクチャ

防御側もAIによる**自律的攻撃**を前提とし、人間の**対応速度**を超えた**自動化されたセキュリティアーキテクチャ**の構築が求められる。