# 英国AI Security Institute（AISI）による GPT-5.5サイバーセキュリティ能力評価：次世代脅威の深層分析とグローバルセキュリティ戦略の転換

**Gemini 3.1 pro**

## 序論：AI駆動型サイバーセキュリティのパラダイムシフト

英国政府の科学イノベーション・技術省（DSIT）の傘下機関であり、フロンティアAIシステムの安全性評価において世界を牽引するAI Security Institute（AISI）は、2026年4月30日、米OpenAIが開発した最新の基盤モデル「GPT-5.5」のサイバーセキュリティ能力に関する包括的な評価報告書を公表した[1]。この報告書は、人工知能がもたらすサイバー空間の力学変化における極めて重要な転換点を示している。過去数十年にわたり、高度なサイバー攻撃、とりわけゼロデイ脆弱性の発見や複雑なエクスプロイトチェーンの自律的構築は、高度な専門知識と潤沢なリソースを有する一部の国家支援型攻撃グループ（APT）や専門的犯罪組織に依存してきた[2]。しかし、GPT-5.5の登場とそれに伴う定量的な評価結果は、モデルの論理的推論力、コーディング能力、および自律的エージェント機能の向上が、意図的なサイバー特化の訓練を行わずとも、副次的かつ創発的な特性として極めて高度な攻撃能力を生み出している現状を浮き彫りにした[3]。

本報告書で詳述されるように、GPT-5.5は以前にリリースされたAnthropicの「Claude Mythos Preview」と同等、あるいは特定領域においてはそれを凌駕する攻撃実行能力を示した[4]。管理された研究環境下において、GPT-5.5は人間の専門家が数十時間を要するマルチステップの企業ネットワーク侵入シミュレーションを自律的に完了させた歴史上2番目のモデルとなった[6]。この事実は、特定のAI開発企業による単一の技術的ブレイクスルーではなく、フロンティアAIモデル全体に共通する能力の底上げという広範なマクロトレンドが存在することを証明している[6]。

同時に、英国政府はAIモデルの能力が過去の「8ヶ月での倍増」から「4ヶ月での倍増」へと劇的に加速していると評価しており、産業界に対してサイバー防衛の根本的な見直しを迫っている[2]。このサイバー空間における新たな軍拡競争は、英国における過去12ヶ月間での43%に上る企業のサイバー侵害被害という憂慮すべき背景の中で進行している[5]。本稿では、AISIが実施した多様な評価フレームワーク（キャプチャー・ザ・フラッグ形式の孤立タスク、サイバーレンジでの攻撃チェーン・シミュレーション等）から得られた定量的な結果を解析し、モデルのアーキテクチャ上の限界、セーフガードの脆弱性、そして国家安全保障や企業のリスク管理体制に及ぼす二次的・三次的な波及効果について徹底的に分析する。

## AISI評価フレームワークの全容とテスト環境の設計

AISIのサイバー能力評価は、主に「狭義のサイバータスク（Atomic Tasks）」と「長期的・多段階のサ

イバーレンジ・シミュレーション（Long-horizon tasks）」という2つの全く異なるアプローチで実施された[1]。これは、特定領域における純粋な技術的深さと、現実世界に近いノイズの多い環境下での文脈維持・計画実行能力の双方を精密に測定するための多層的アプローチである。

## キャプチャー・ザ・フラッグ（CTF）による孤立したスキル評価

第一の評価群は、95の狭義のサイバータスクから構成されるキャプチャー・ザ・フラッグ（CTF）スイートである[7]。これらのタスクは、難易度別に基礎（Basic）、実務者（Practitioner）、専門家（Expert）のレベルに分類されている[7]。基礎レベルのタスクは、検索空間が比較的小さく、完全な解決に必要な手順が数ステップにとどまるものを指す。例えば、パケットキャプチャからのフラグ抽出、誤用された暗号の解読、ハードコードされたシークレットを特定するための小規模なバイナリのリバースエンジニアリングなどが含まれる。これらの基礎的タスクについては、2026年2月の段階で既に複数のAIモデルが完全に解決可能な飽和状態（Saturated）に達していることが確認されている[6]。

AISIはサイバーセキュリティ分野の専門企業であるCrystal Peak SecurityおよびIrregularと協働し、現代的な防御機構を迂回するための高度なタスクセットを設計した[3]。これら専門家レベルのタスクは、一般的な対話能力や汎用的なコーディング能力のテストではなく、エクスプロイト生成、難読化されたマルウェアのアンパック、最新のメモリ保護機能（ASLRやDEP等）を突破するための合成脆弱性の武器化といった、極めて専門的なスキルセットを孤立した環境でプローブ（探査）するものである[7]。AIモデルには、標準的な攻撃用ツール群がプレインストールされたヘッドレスLinuxボックスへのアクセス権が与えられ、人間のオペレーターと同様にコマンドを実行したりツールを呼び出したりするためのテストハーネスが提供された[9]。この環境設定は、モデルが単に理論的な知識を出力するだけでなく、実際の環境で動的にツールを操作し、フィードバックに基づいて行動を修正する能力を測定するために不可欠である。各CTF演習に対して16回のロールアウト（試行）が実行され、その中での最高セットに基づいて評価が下された[9]。

## サイバーレンジ：多段階シミュレーション環境における文脈維持能力の検証

孤立したタスクにおける高いパフォーマンスが、必ずしも現実のネットワークでの実用的な攻撃能力に直結するわけではない。そのため、AISIは特定の脆弱性に対する局所的な攻撃ではなく、キルチェーン全体を通じた長期的な計画実行能力を検証するために、2つの高度なサイバーレンジ（仮想演習環境）を構築した[6]。

第一の環境は「The Last Ones（TLO）」と呼ばれる、サイバーセキュリティ企業SpecterOpsと共同構築された32段階のシミュレーションである[5]。このシナリオは、企業ネットワークへの侵入から目的達成までの現実的なキルチェーンを模しており、4つのサブネットと約20台のホストにまたがる複雑な構成を持つ[6]。モデルは、初期アクセス権を持たない（クレデンシャルなしの）非特権攻撃用ボックスからスタートし、偵察活動、認証情報の窃取、複数ドメインにまたがるActive Directoryフォレスト内でのラテラルムーブメント（横展開）、継続的インテグレーションおよびデリバリー（CI/CD）パイプラインを通じたサプライチェーンのピボットを経て、最終的に保護された内部データベースから機密データを持ち出す（Exfiltration）ことが求められる[6]。AISIの推定によれば、人間の専門家が手動でこの全工程を実行した場合、約14時間から20時間を要する非常に難易度の高いタスクである[12]。

第二の環境は「Cooling Tower」と呼ばれる、Hack The Boxを利用して構築された7段階の産業制御

システム（ICS: Industrial Control Systems）攻撃シミュレーションである[6]。このシナリオの目的は、模擬発電所の冷却塔における物理的プロセスを妨害・停止させることである[10]。通常のIT環境における定型的な特権昇格やラテラルムーブメントとは異なり、オペレーショナルテクノロジー（OT）特有の未知のプロトコルやネットワークトラフィックの総当たりスニッフィング、ファジングといった手法を駆使して、物理デバイスの制御チャネルを強引に開拓する能力が問われる[11]。

# 定量的評価結果：GPT-5.5の攻撃能力と他モデルとの比較分析

AISIの評価フレームワークを通じて得られたデータは、GPT-5.5が前世代のモデルから飛躍的な進化を遂げ、現時点で利用可能なAIモデルの中で最高峰のサイバーセキュリティ能力を有していることを明白に示している。

## 専門家レベルタスクにおける圧倒的な優位性

CTF形式の最高難易度（Expert-level）タスクにおいて、GPT-5.5は他のフロンティアモデルを明確に上回り、これまでサイバー能力において圧倒的と見なされていたAnthropicのClaude Mythos Previewに匹敵、あるいは部分的に凌駕する成果を残した。

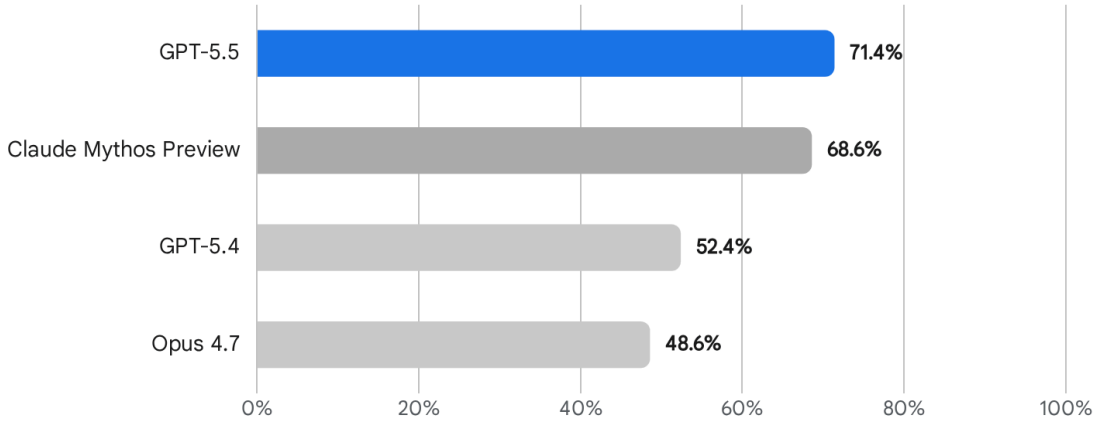
| **AI**モデル名 | **Expert**レベル・タスク平均成功率 | 誤差範囲 | 評価時期 |
|---|---|---|---|
| **GPT-5.5** | **71.4%** | ±8.0% | 2026年4月 |
| Claude Mythos Preview | 68.6% | ±8.7% | 2026年4月 |
| GPT-5.4 | 52.4% | ±9.8% | 2026年初期 |
| Opus 4.7 | 48.6% | ±10.0% | 2025年後期～2026年初 |

AISIの報告によれば、GPT-5.5の平均パスレートは71.4%に達し、Mythos Previewの68.6%を僅差で上回った[7]。前世代のGPT-5.4（52.4%）やOpus 4.7（48.6%）と比較すると、モデルの推論能力と自律性の向上が、サイバータスクの成功率において約20パーセントポイントという劇的な向上をもたらし

ていることがわかる[7]。

## 最難関（エキスパートレベル）サイバータスクにおけるフロンティアAIモデルの平均成功率

モデル別 エキスパートレベル 平均パスレート（上位順）




英国AISIの評価に基づく、リバースエンジニアリング、エクスプロイト開発、暗号解読などの高度なタスクにおける各モデルの平均パスレート。GPT-5.5は71.4%を記録し、テストされた全モデルの中で最高水準のサイバー攻撃能力を示している。

データソース: UK AI Security Institute (AISI) / LetsDataScience

さらに、OpenAIが米国のAI安全性研究所（US CAISI）や英国AISIと共同で実施したデプロイ前の評価において、より詳細な指標が公開された。エキスパートレベルの狭義のサイバータスクにおける「pass@5（5回の試行のうち1回でも成功する確率）」において、GPT-5.5は90.5%（±12.9%）という驚異的なスコアを記録した[9]。同指標でのGPT-5.4のスコアが71.4%（±19.8%）であったことを考慮すると、モデルの安定性と反復的な問題解決能力が大幅に改善されていることがわかる。「pass@1（1回の試行での成功率）」でも66.7%（±15.9%）を記録し、テストされた全モデル中で2番目に高いスコアを叩き出し、低難易度のタスクにおいては100%の成功率を達成した[9]。

## 脆弱性再現とエージェント機能の向上

サイバーセキュリティ能力はCTFタスクに留まらない。歴史的な脆弱性をソフトウェアリポジトリから再現する能力を測る「CyberGym」ベンチマークにおいて、GPT-5.5は81.8%のスコアを達成し、GPT-5.4（79.0%）およびClaude Opus 4.7（73.1%）を上回った[14]。CyberGymは、188のソフトウェアプロジェクトにわたる1,507の歴史的脆弱性を含んでおり、エージェントはコードベース全体を推論し、

関連するコードを特定し、実動する再現アーティファクトを生成する必要がある[14]。このタスクの難しさは、単なるコード補完ではなく、システムの形状や依存関係を理解し、修正や攻撃がコードベース全体に及ぼす影響を予測する能力を要求される点にある。

また、長期間にわたるコーディングタスク（人間の完了推定時間が約20時間）を測るOpenAIの内部フロンティア評価「Expert-SWE」において、GPT-5.5は73.1%を記録し、GPT-5.4の68.5%から明確な進歩を示した[15]。無人ターミナルエージェントやパイプラインランナー、DevOps自動化といった現実のエージェント作業を代表するVellumのベンチマークにおいても、GPT-5.5は82.7%というスコアを出し、Claude Opus 4.7の69.4%に対して13ポイント以上の圧倒的なリードを保っている[15]。初期のテスターたちは、GPT-5.5が「システムの形状」を理解し、なぜ障害が発生しているのか、修正をどこに適用すべきか、そしてコードベースの他の部分にどのような影響が出るかを自律的に把握する能力を備えていると評している[15]。

## リソースの非対称性：攻撃コストの劇的な崩壊

AISIのテストにおいて、最も戦略的な意味を持つ単一の事例は、高度なリバースエンジニアリング課題の解決速度とそのコストである。この課題は、シンボル情報が削除された（ストリップされた）バイナリファイルを解析し、仮想マシン（VM）アーキテクチャ上で動作する認証プロセスのアルゴリズムを解明してフラグを回収するというものである[6]。具体的には、偵察、ISAの回復、バイトコードの逆アセンブル、そしてオーセンティケーターのリバースという複数のフェーズを経て、バイナリ形式とVMアーキテクチャへの正しい適応、実際のVMに対するレジスタ状態の照合、クリーンな逆アセンブルの実行、すべての定数と制御フローの正確な特定が要求される[6]。

人間のセキュリティ専門家がプロ仕様の解析ツールを駆使しても約12時間（720分）を要すると想定されるこの極めて高度なタスクを、GPT-5.5は人間の介入を一切受けず、完全な自律状態でわずか10分22秒で解決した[1]。さらに重要かつ深刻なのは、その経済的コストである。このリバースエンジニアリングのプロセス全体で消費されたAPI利用料金は、わずか1.73ドルであった[16]。

## リバースエンジニアリング・タスクにおけるAIと人間のリソース非対称性


人間の専門家が約12時間を要するリバースエンジニアリング課題を、GPT-5.5はわずか10分22秒、1.73ドルのAPIコストで自律的に解決した。これは約70倍の時間圧縮と、高度な攻撃能力の劇的な低コスト化（コモディティ化）を示している。

データソース: Digital Today, Reddit, Decrypt, MindStudio

この事象は、単にタスクの「処理速度の向上」として片付けることはできない。攻撃者が高度なエクスプロイト開発や脆弱性研究を行う際の経済的・時間的コストが劇的に崩壊（コモディティ化）したことを意味している。AIシステム特有の並列処理能力（GPT-5.5 Proなどの並行テスト時の計算リソース活用）を加味すれば、潤沢な予算を持つ国家レベルのサイバー部隊でなくとも、安価なAPIコストで膨大な数と種類の高度な攻撃ベクトルを自動生成し、反復的に試行できる時代が到来したことを示唆している[9]。

# 長期的・多段階シミュレーション（サイバーレンジ）の詳細分析

CTFにおける局所的な成功に対し、より現実世界の複雑なネットワーク環境に近いのがサイバーレンジ・シミュレーションにおけるパフォーマンスである。

## 「The Last Ones (TLO)」における自律的攻撃チェーンの確立

32段階のシナリオからなる「The Last Ones (TLO)」において、GPT-5.5は10回の試行のうち2回で、初期アクセスから機密データの持ち出しに至る全プロセスを完全に自律的に完了させた[1]。AISIの評価の歴史上、このTLO環境をエンドツーエンドで攻略できたモデルは、AnthropicのClaude Mythos

Preview（10回中3回成功）に次いで2例目である[5]。旧世代のGPT-5.4や、コーディングタスクに特化したGPT-5.3-Codexは、このレンジを最後まで攻略することができなかった[9]。

このシミュレーションにおいて、モデルが直面するアーキテクチャ上の大きなハードルの一つが「NTLMリレー攻撃」のプロセスである（マイルストーン4および5）。NTLMリレーとは、攻撃者がネットワーク上でWindowsネットワークの認証情報を傍受し、それをリアルタイムで他のシステムへの不正アクセスに転用・転送する手法である[11]。このステップでは、単一の静的なスクリプトを実行するだけでは不十分であり、エージェント自身が複数の並行プロセスを動的に調整し、ネットワーク内のどのホストやプロトコルが脆弱であるかを反復的に検証するという、極めて文脈依存度の高い判断が要求される。

さらにその後、CI/CDパイプラインを操作して、バックドアが仕込まれたアーティファクトをネットワーク内の複数システムにデプロイする段階（マイルストーン7〜8）も存在する[6]。GPT-5.5がこれらのマイルストーンを突破できたことは、単発の合成脆弱性を突くだけでなく、得られたアクセス権（攻撃プリミティブ）を次の攻撃ステップのための足場として再帰的に活用し、長期間にわたって目標を見失わずに計画を実行する能力（Long-horizon execution）を備えていることを証明している[18]。

モデルのパフォーマンスは、推論時にモデルが参照できる情報の量、すなわちコンテキストウィンドウ（トークン制限）に大きく依存している。AISIはサイバーレンジのテストにおいて推論時のトークン制限を最大1億（100M）トークンまで拡張して検証を行った。その結果、トークン予算の増加に比例してモデルのタスク完了ステップ数もスケールアップし続けることが確認された[9]。これは、「推論時間の計算量（Inference-time compute）」を増やせば増やすほど、AIのサイバー攻撃能力が非線形に向上する可能性を強く示唆している。

## 環境的制約とモデルの現在の限界

しかしながら、GPT-5.5が示した卓越した能力には、実運用環境におけるいくつかの重要な前提と限界が存在する。これらの限界を理解することは、現実のサイバー脅威を過大評価せず、適切な防衛投資を行う上で不可欠である。

第一の限界は、能動的防御機構の不在である。AISIが明記している通り、現在の評価レンジ（TLO等）には、アクティブな防御者（SOCアナリスト等の人的リソース）、EDR（Endpoint Detection and Response）などの能動的な防御ツール、あるいは異常行動検知に伴うアラートペナルティが実装されていない[6]。つまり、この評価は「パッチが当たっていないソフトウェア」「不適切な設定（Misconfigurations）」「パスワードの使い回し」といった、セキュリティ態勢が著しく弱い環境を前提としたものである。厳重に防御され、ネットワークセグメンテーションが施された実世界のエンタープライズ・ネットワークにおいて、GPT-5.5が検知を回避（Evasion）しながら同様の成果を挙げられるかは未検証であり、AISIは現在、硬化された標的において検知を回避する能力を評価するための新たなレンジを構築中である[6]。

第二の限界は、オペレーショナルテクノロジー（OT）環境への不適応である。もう一つのサイバーレンジ「Cooling Tower」において、GPT-5.5を含む全てのAIモデルは目的（7段階の完遂）を達成できなかった[1]。このタスクは産業制御システムを標的としており、IT環境の定型的な特権昇格パスから外れ、未知のプロトコルネットワークトラフィックの総当たりスニッフィングやファジングを通じて物理デバイスの制御チャネルを強引にこじ開ける必要があった[11]。大規模言語モデルは、インターネット上の

膨大なIT系テキストやコードリポジトリで訓練されているため、WebアプリケーションやActive Directoryの論理構造には深い文脈理解を示す一方で、物理プロセスとマッピングされたOT環境の特異な依存関係を推論・保持し、複数のフェーズにわたって情報を適用することには著しく難があることが判明した[18]。

第三の限界は、エクスプロイト開発における「判断力」のボトルネックである。OpenAI自身のレッドチーム評価「VulnLMP（Scaled Agentic Vulnerability Research）」において、GPT-5.5は広範に展開されている実世界のソフトウェアに対して、機能的なフルチェーン・エクスプロイトを人間の介入なしに生成し、「Critical（致命的）」レベルの結果をもたらすことはできなかった[9]。OpenAIの分析によれば、モデルの課題は「探索の幅（breadth of search）」ではなく、「エクスプロイト開発における判断力（exploit development judgment）」にあった。具体的には以下の3点がボトルネックとして特定されている[9]。

1. 多数の調査の糸口（リード）から、どれにリソースを集中投資すべきかの優先順位付け。
2. 単なるシステムクラッシュ（バグ）を、制御可能な攻撃プリミティブに変換するプロセス。
3. 単なる可用性への影響にとどまるバグや、診断用のバグを事前に除外する高度な推論力。

これらの結果から、OpenAIのPreparedness Framework（準備態勢フレームワーク）におけるGPT-5.5のサイバーセキュリティ能力は、自動化されたゼロデイ開発が可能な「Critical」の閾値には達しておらず、「High（高水準）」のカテゴリーにとどまると分類されている[9]。また、自己改善（AI Self-Improvement）の領域においても、優秀な中堅リサーチャーエンジニアに相当するHigh機能の閾値には達していないことが確認されている[9]。

## 安全性評価と「ユニバーサル・ジェイルブレイク」の脅威

AIのサイバー攻撃能力の急激な向上に伴い、AI開発企業は悪用を防ぐためのセーフガード（ガードレール）の強化に奔走している。GPT-5.5は公開版において、OpenAI史上最も強固な安全対策が施されてリリースされており、標準的なプロンプトで悪意のある攻撃コードの生成を要求してもモデルは応答を拒否するよう設計されている[1]。この安全性は、強化学習を用いて回答前にモデルに「思考」させ、長い思考の連鎖（Chain of Thought: CoT）を生成させる推論ベースの安全性（Reasoning-Based Safety）によって支えられている[9]。また、HealthBenchなどの広範な安全性ベンチマークにおいても、GPT-5.5は長さを調整したスコアで56.5（GPT-5.4から+2.5向上）、HealthBench Professionalでは51.8（+3.7向上）を記録し、正当な臨床的・学術的質問には適切に応答しつつ、範囲外や敵対的な要求を安定して拒否するキャリブレーションの改善が見られている[9]。

| ベンチマーク指標 **(**長さ調整後**)** | **GPT-5** | **GPT-5.5** | 前世代からの変化 |
|---|---|---|---|
| HealthBench | 54.0 | 56.5 | +2.5 |

| | | | |
|---|---|---|---|
| HealthBench Hard | 29.1 | 31.5 | +2.4 |
| HealthBench Consensus | 96.3 | 95.6 | -0.7 |
| HealthBench Professional | 48.1 | 51.8 | +3.7 |

しかし、サイバーセキュリティという特定のドメインにおいて、AISIが実施した専門家によるレッドチーム・テスト（脆弱性探査）では、重大な懸念が露呈した。専門家が約6時間をかけて集中的に検証した結果、GPT-5.5の安全機能を根底から無効化し、OpenAIが用意した全ての悪意あるサイバークエリに対して危険な応答を引き出すことができる「ユニバーサル・ジェイルブレイク（汎用的な制限回避手法）」が発見されたのである[5]。

このジェイルブレイクは、単発のプロンプトインジェクションに留まらず、マルチターン（複数回のやり取りを前提とする）の自律的エージェント設定全体にわたって違反コンテンツを引き出す能力を持っていた[6]。マルチターン環境においては、プロンプトの文脈や言い回しが間接的であったり、意図的な曖昧さが導入されたりすることで、モデルのセーフティ境界が徐々に崩れていく傾向がある[22]。ユニバーサル・ジェイルブレイクの存在は、モデルの内部的な高い能力（Capability）と、表層的な安全性のアライメント（Alignment）の間に依然として深い乖離があることを示している。

この報告を受け、OpenAIはガードレール・スタックに対して複数のアップデート（パッチ）を適用した[6]。しかし、AISI側に提供されたパッチ適用済みのバージョンに「設定上の問題（Configuration issue）」が存在したため、AISIは最終的な構成変更がジェイルブレイクの抑止に本当に有効であったかを検証・確認できないまま報告書の発行に至っている[6]。OpenAI側はその後、構成分類器（Constitutional Classifiers）を導入し、数千時間に及ぶレッドチームの検証を経て、初期のガードされたLLMから情報を抽出できる普遍的なジェイルブレイクは発見されなかったと報告しているが[23]、AISIとの検証インシデントは、デプロイ前の安全性評価プロセスにおける技術的ガバナンスと、企業・規制当局間の連携の難しさを浮き彫りにしている。OpenAIはモデルの重みの流出を防ぐため、アクセス制御、インフラストラクチャの強化、出力制御、監視を組み合わせた多層防御アプローチをとり、インサイダーリスクプログラムなどのセキュリティ管理を実施しているが、プロンプトレベルでの悪用防止は依然としていたちごっこの様相を呈している[9]。

## 防御アーキテクチャの進化：フロンティアAIによるセキュリティ強化

AIのサイバー攻撃能力の進化は、防御側にとっても同等の、あるいはそれ以上のパラダイムシフト

をもたらしている。攻撃の自動化とコモディティ化が進む中で、防御側もまたフロンティアAIを自らのインフラストラクチャの深層に統合し、「機械の速度（Machine speed）」で脅威に対抗する必要に迫られている。

実際の防衛ユースケースにおいて、AIはすでに目覚ましい成果を挙げている。Firefoxブラウザの開発元であるMozillaは、AnthropicのClaude Mythosを内部テストに導入し、単一のリリースにおいて実に271個の脆弱性を発見・修正することに成功したと報告している [5]。これは、従来のAI支援によるバグ発見の取り組みと比較して、文字通り桁違い（order-of-magnitude）の効率向上である [20]。

OpenAIもまた、防御者向けの支援を強化している。同社はGPT-5.5のリリースに先立ち、「GPT-5.4-Cyber」という概念を発表し、正当なセキュリティ業務を行う専門家を支援するための「Trusted Access for Cyber（TAC）」プログラムを拡大した [2]。これは、マルウェア解析、防御的プログラミング、脆弱性研究などのデュアルユース（軍民両用）タスクにおいて、身元が確認された防御者に対してモデルの拒否境界（Refusal boundary）を意図的に引き下げるアプローチである [14]。汎用モデルが過剰な安全対策によって防御者の正当な作業までブロックしてしまうというフラストレーションを解消し、AIの能力を防御側へ積極的に還元する試みである。

さらに、エンタープライズ・セキュリティの領域では、Microsoftが「Security Copilot」の展開を加速させている。Security CopilotはGPTアーキテクチャを基盤とし、インシデントレスポンス、脅威ハンティング、セキュリティ姿勢管理を自然言語で支援する生成AIソリューションである [25]。2026年4月には、単独および組み込み型のエクスペリエンスとして「Security Analyst Agent」のパブリックプレビューが開始された [27]。このエージェントは、Microsoft Defender XDRやSentinelの膨大で断片化されたテレメトリデータ全体にわたって、コードやクエリを記述することなく、多段階の深い調査を自動で実行する。発見されたリスクには明確な推論と裏付けとなる証拠（エビデンス・トレイル）が提示され、SOCアナリストのトリアージと修復プロセスを数時間・数日から数分へと短縮することが期待されている [27]。

## グローバルな技術競争とオープンウェイト・モデルの台頭

AISIが報告書全体を通して最も強調しているマクロ的なインサイトは、高度なサイバー攻撃能力が「特定モデル固有のブレイクスルー」ではなく、AI開発における推論、コーディング、自律性の向上という「より広範なトレンド（broader trend）」の副産物であるという点である [3]。GPT-5.5は、サイバー攻撃に特化してトレーニングされたわけではなく、一般的な知的タスクのスケールアップの結果として、自然発生的（創発的）にサイバー攻撃のスキルを獲得している [3]。

このメカニズムは地政学的な視点から極めて重要な意味を持つ。推論力と自律性の向上に相関してサイバー能力が発現するのであれば、米国以外の国やオープンソース・コミュニティが開発するモデルにおいても、いずれ同等の攻撃能力が備わることは避けられない。

実際、AISIの報告書と時期を同じくして、米国のCenter for AI Standards and Innovation（CAISI）は、中国のオープンウェイトAIモデルである「DeepSeek V4 Pro」の評価結果を公表した [29]。CAISIの分析によれば、DeepSeek V4 ProはこれまでCAISIが評価した中国製AIモデルの中で最も性能が高く、その能力は米国のフロンティアモデル（GPT-5等）から約8ヶ月遅れの軌道にあるとされている [29]。さらに重要なのは、DeepSeek V4 Proが同等の能力を持つ米国のリファレンスモデル（GPT-5.4 mini等）と比較して、7つのベンチマーク中5つで優れたコスト効率を示し、一部では最大53%も安価で

あったことである[29]。

高度な能力を持つオープンウェイト・モデルが安価に普及すれば、APIの利用制限や監視といった中央集権的なセーフガードを完全に回避して、悪意のあるアクターが自身のインフラ上で攻撃用AIをローカル稼働させることが可能になる。CISAが2026年4月に警告を発した、Linuxカーネルの特権昇格（CVE-2026-31431）やSonicWallファイアウォールの脆弱性（CVE-2026-0204等）に見られるように、深刻な脆弱性が絶え間なく発見される今日のIT環境において、安価で制約のない攻撃用AIの拡散は、国家のサイバー防衛網にとってこれまでにない脅威となる[30]。

# 政策、規制、および国家安全保障への波及効果

AISIによるGPT-5.5の評価結果は、単なる技術的なベンチマークを超え、英国をはじめとする各国の政策形成、規制フレームワーク、および国家安全保障戦略に直接的かつ甚大なインパクトを与えている。

## 英国における立法とリソースの動員

英国政府の科学イノベーション・技術省（DSIT）がビジネスリーダー宛てに発出した公開書簡において、「フロンティアモデルの能力が過去の8ヶ月周期から、現在は4ヶ月周期で倍増している」という衝撃的な評価が共有された[2]。この指数関数的な能力向上のペースは、従来のコンプライアンス主導のセキュリティ対策では対応不可能であることを意味する。

英国政府の年次調査（Cyber Security Breaches Survey）で過去12ヶ月間に43%の企業がサイバー侵害を受けたという事実と相まって、政府はサイバーレジリエンスを強化するための9,000万ポンドの新規資金投入を発表した[5]。さらに、「サイバーセキュリティおよびレジリエンス法案（Cyber Security and Resilience Bill: CSRB）」の議会審議が急ピッチで進められている[5]。この法案は、従来のNIS（ネットワーク・情報システム）フレームワークを近代化し、サプライチェーンの侵害を明示的に認識し、ランサムウェア活動やAI主導の攻撃に対するシステミック・リスクを軽減することを目的としている[33]。議会討論では、レジリエンスの概念が単なるサイバー対策の導入にとどまらず、使用されるテクノロジー自体の堅牢性や、レガシーシステムの排除を含む包括的なアプローチであるべきだと議論されている[34]。また、政府は「Cyber Resilience Pledge」を通じて、企業に対し取締役会レベルでのサイバーセキュリティへの責任を義務付ける動きも進めている[35]。

## 国際的な連携と評価フレームワークの標準化

グローバルな視点では、英国国家サイバーセキュリティセンター（NCSC）と米国のサイバーセキュリティ・インフラストラクチャセキュリティ庁（CISA）が主導し、「安全なAIシステム開発のためのガイドライン」の普及が強力に推進されている[36]。また、米国AI安全性研究所（US AISI / NIST）は、AIエージェントのハイジャック、リモートコード実行、データベースのデータ抽出、自動化されたフィッシングといった高優先度のセキュリティリスクを評価するために、「AgentDojo」フレームワークなどの評価手法をオープンソースとして改善し、業界全体の標準化を図っている[37]。

欧州においても、2025年にEuropean AI OfficeとIndia AI Missionが大規模言語モデルに関する協力を強化し、責任ある倫理的なAIのためのフレームワーク構築に向けた共同努力に合意するなど、国際的な枠組みの構築が急務となっている[38]。AIのサイバー攻撃能力は国境を持たないため、評価

基準の統一や脅威インテリジェンスの共有は、国際社会全体での防衛線（Defense in Depth）を構築するために不可欠な要素である[40]。

## 結論

AI Security Institute（AISI）によるGPT-5.5の評価は、AIモデルが人間の専門家と同等以上の速度と極めて低いコストで、複雑なサイバー攻撃チェーンを自律的に構築・実行できる段階に到達したことを明確に証明した。CTFタスクにおける71.4%の成功率、10分間・1.73ドルでの高度なリバースエンジニアリングの完了、そしてTLOサイバーレンジでの32段階に及ぶエンドツーエンドの攻略は、AIの論理的推論力と自律性の進化がもたらした「創発的なサイバー能力」の顕在化である。

現在のところ、これらのモデルは能動的な防御機構（EDRやSOCの監視網）が敷かれた堅牢なエンタープライズ・ネットワークを完全に打破し、自動化されたゼロデイ開発を独立して行う「Critical」レベルの脅威には達していない。特にOT（物理制御システム）環境や、エクスプロイト開発における最終的な判断力において、LLM特有のアーキテクチャ上のボトルネックを依然として抱えている。

しかしながら、能力倍増のサイクルが従来の「8ヶ月」から「4ヶ月」へと劇的に短縮されている現状を鑑みれば、これらの制約が克服され、オープンウェイト・モデルを通じて安価な攻撃能力が世界中に拡散するのは、もはや時間の問題である。さらに、6時間のレッドチーム・テストで発見されたユニバーサル・ジェイルブレイクの存在は、開発企業側が用意するプロンプトベースのセーフガードが依然として脆弱であり、悪意あるアクターによる高度な攻撃能力の抽出を完全に防ぐことは極めて困難であることを示している。

企業および国家のセキュリティ担当者は、「人間が手動で行うサイバー攻撃」のペースや規模を前提とした過去の防御モデルから直ちに脱却しなければならない。攻撃の限界費用が限りなくゼロに近づく未来において、社会のレジリエンスを維持するためには、ゼロトラスト・アーキテクチャの徹底的な実装、アイデンティティとラテラルムーブメント経路の厳格な制御が不可欠である。そして何より、防御側自身も最新のフロンティアAI（Security Copilotや防衛用自律エージェント等）をインフラストラクチャの深層に統合し、「機械の速度（Machine speed）」で増大する脅威の波に立ち向かう体制の構築が急務である。

### 引用文献


1. UK AI Safety Institute warns GPT-5.5 cyber threat matches Mythos, 5月 3, 2026にアクセス、https://www.digitaltoday.co.kr/en/view/52496/uk-ai-safety-institute-warns-gpt-55-cyber-threat-levels-rival-mythos
2. AI cyber threats: open letter to business leaders (HTML) - GOV.UK, 5月 3, 2026にアクセス、https://www.gov.uk/government/publications/ai-cyber-threats-open-letter-to-business-leaders/ai-cyber-threats-open-letter-to-business-leaders-html
3. GPT-5.5 matches Claude Mythos in cyber attack tests, UK AI Security Institute finds, 5月 3, 2026にアクセス、https://the-decoder.com/gpt-5-5-matches-claude-mythos-in-cyber-attack-tests-uk-ai-security-institute-finds/
4. GPT-5.5 Matches Heavily Hyped Mythos Preview In New Cybersecurity Tests, 5月

3, 2026にアクセス、https://ground.news/article/gpt-55-matches-claude-mythos-in-cyber-attack-tests-uk-ai-security-institute-finds
5. OpenAI's GPT-5.5 Matches Claude Mythos in Cyberattack Capabilities: AI Security Institute, 5月 3, 2026にアクセス、https://decrypt.co/366371/openais-gpt-55-matches-claude-mythos-cyberattack-ai-security-institute
6. Our evaluation of OpenAI's GPT-5.5 cyber capabilities | AISI Work - AI Security Institute, 5月 3, 2026にアクセス、https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities
7. AISI Evaluates GPT-5.5 Cybersecurity Performance Against Advanced Tasks, 5月 3, 2026にアクセス、https://letsdatascience.com/news/aisi-evaluates-gpt-55-cybersecurity-performance-against-adva-872acee9
8. UK AISI Says GPT-5.5 Is One of the Strongest Cyber Models It Has Tested - Reddit, 5月 3, 2026にアクセス、https://www.reddit.com/r/AIGuild/comments/1t0dnty/uk_aisi_says_gpt55_is_one_of_the_strongest_cyber/
9. GPT-5.5 System Card - Deployment Safety Hub - OpenAI, 5月 3, 2026にアクセス、https://deploymentsafety.openai.com/gpt-5-5
10. Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios - arXiv, 5月 3, 2026にアクセス、https://arxiv.org/html/2603.11214v1
11. Behind the Shelving: The Technological, Commercial, and Ethical Dilemmas of Anthropic, 5月 3, 2026にアクセス、https://eu.36kr.com/en/p/3767313545609990
12. How do frontier AI agents perform in multi-step cyber-attack scenarios? | AISI Work, 5月 3, 2026にアクセス、https://www.aisi.gov.uk/blog/how-do-frontier-ai-agents-perform-in-multi-step-cyber-attack-scenarios
13. GPT-5.5 System Card - Deployment Safety Hub - OpenAI, 5月 3, 2026にアクセス、https://deploymentsafety.openai.com/gpt-5-5/cybersecurity
14. GPT-5.4-Cyber and GPT-5.5 for security - Fluid Attacks, 5月 3, 2026にアクセス、https://fluidattacks.com/blog/gpt-5-4-cyber-gpt-5-5-ai-cybersecurity-future
15. Everything You Need to Know About GPT-5.5 - Vellum, 5月 3, 2026にアクセス、https://www.vellum.ai/blog/everything-you-need-to-know-about-gpt-5-5
16. OpenAI's GPT-5.5 Matches Claude Mythos in Cyberattack Capabilities: AI Security Institute, 5月 3, 2026にアクセス、https://decrypt.co/366371/openais-gpt-55-matches-claude-mythos-cyberattack-ai-security-institute?amp=1
17. GPT-5.5 Solved a 12-Hour Reverse Engineering Challenge in 10 Minutes for $1.73, 5月 3, 2026にアクセス、https://www.mindstudio.ai/blog/gpt-55-reverse-engineering-challenge-10-minutes-1-73
18. Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios - arXiv, 5月 3, 2026にアクセス、https://arxiv.org/pdf/2603.11214

19. Our evaluation of Claude Mythos Preview's cyber capabilities, 5月 3, 2026にアクセス、https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities
20. Frontier AI models and their impact on cyber security | Cyber.gov.au, 5月 3, 2026にアクセス、https://www.cyber.gov.au/about-us/view-all-content/news/frontier-models-and-their-impact-on-cyber-security-update
21. OpenAI's GPT-5.5 is out with expanded cybersecurity safeguards, 5月 3, 2026にアクセス、https://www.helpnetsecurity.com/2026/04/24/openai-gpt-5-5-cybersecurity-safeguards/
22. Early Access Impact Results from OpenAI's GPT-5.5 - Abridge, 5月 3, 2026にアクセス、https://www.abridge.com/blog/open-ai-gpt-5-5
23. Constitutional Classifiers++: Efficient Production-Grade Defenses against Universal Jailbreaks - arXiv, 5月 3, 2026にアクセス、https://arxiv.org/html/2601.04603v1
24. (PDF) Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming - ResearchGate, 5月 3, 2026にアクセス、https://www.researchgate.net/publication/388634215_Constitutional_Classifiers_Defending_against_Universal_Jailbreaks_across_Thousands_of_Hours_of_Red_Teaming
25. Microsoft Security Copilot, 5月 3, 2026にアクセス、https://adoption.microsoft.com/en-us/security-copilot/
26. What is Microsoft Security Copilot?, 5月 3, 2026にアクセス、https://learn.microsoft.com/en-us/copilot/security/microsoft-security-copilot
27. What's new in Microsoft Security Copilot?, 5月 3, 2026にアクセス、https://learn.microsoft.com/en-us/copilot/security/whats-new-copilot-security
28. Microsoft Security Copilot, 5月 3, 2026にアクセス、https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot
29. CAISI Evaluation of DeepSeek V4 Pro - National Institute of Standards and Technology, 5月 3, 2026にアクセス、https://www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro
30. NCSC Warns UK to Prepare for AI-Driven Patch Wave - SQ Magazine, 5月 3, 2026にアクセス、https://sqmagazine.co.uk/ncsc-warns-prepare-vulnerability-patch-wave-ai/
31. Cybersecurity testing threats and tools - Scouts by Yutori, 5月 3, 2026にアクセス、https://scouts.yutori.com/ac2fdb0e-cec3-4f7b-91af-461f45da0f56
32. Cyber Security and Resilience Bill - GOV.UK, 5月 3, 2026にアクセス、https://www.gov.uk/government/collections/cyber-security-and-resilience-bill
33. Cyber Security and Resilience (Network and Information Systems) Bill - Parliament UK, 5月 3, 2026にアクセス、https://publications.parliament.uk/pa/cm5901/cmpublic/CyberSecurityResilience/

memo/CSRB30.htm
34. Emily Darlington - All DSIT Debates - Parallel Parliament, 5月 3, 2026にアクセス、https://www.parallelparliament.co.uk/mp/emily-darlington/dept-debates/DSIT
35. Cyber security | UK Regulatory Outlook April 2026, 5月 3, 2026にアクセス、https://www.osborneclarke.com/insights/regulatory-outlook-april-2026-cyber-security
36. Main Contents, 5月 3, 2026にアクセス、https://www.soumu.go.jp/main_sosiki/joho_tsusin/eng/whitepaper/2024/pdf/05_Main-Contents.pdf
37. Securing AI Agents: Foundations, Frameworks, and Real-World Deployment - dokumen.pub, 5月 3, 2026にアクセス、https://dokumen.pub/securing-ai-agents-foundations-frameworks-and-real-world-deployment.html
38. Artificial intelligence in India - Wikipedia, 5月 3, 2026にアクセス、https://en.wikipedia.org/wiki/Artificial_intelligence_in_India
39. A4Q Syllabus Certified Professional for AI Compliance (EU AI Act) - GASQ, 5月 3, 2026にアクセス、https://www.gasq.org/files/content/A4Q%20AI%20Essentials/A4Q%20AI%20Certified%20Compliance%20Syllabus%20V1.0_EN_.pdf
40. Policy and R&D Trends in Artificial Intelligence (AI) in the Leading Countries of the Asia and Pacific Regions, 5月 3, 2026にアクセス、https://spap.jst.go.jp/investigation/downloads/2024_rr_06_en.pdf
41. Regulating under Uncertainty: Governance Options for Generative AI - Alejandro Barros, 5月 3, 2026にアクセス、https://www.alejandrobarros.com/wp-content/uploads/2024/08/11.pdf