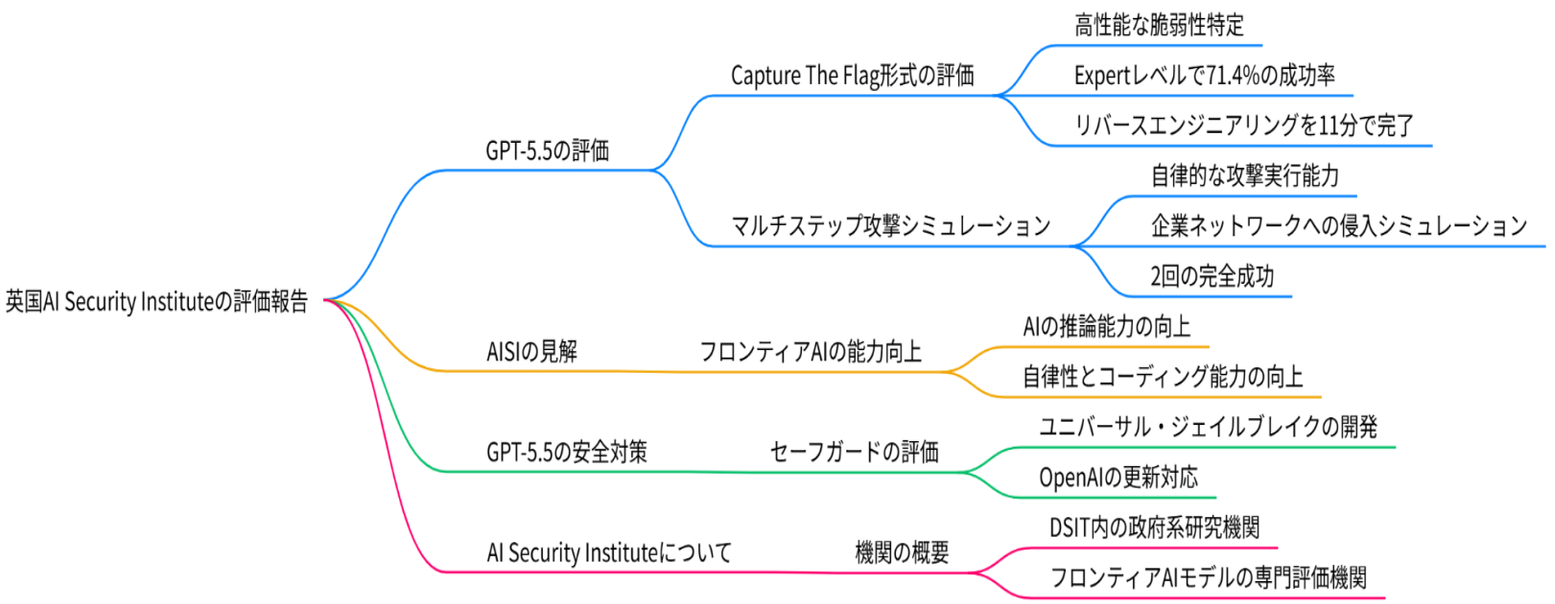
英国政府の AI Security Institute（AISI）が「GPT-5.5」のサイバーセキュリティ能力に関する包括的な評価報告書を公表

Felo AI



英国の AI Security Institute（AISI）は 2026 年 4 月 30 日、OpenAI が開発した AI モデル「GPT-5.5」のサイバーセキュリティ能力に関する評価報告書を公表しました[1, 5, 8]。評価の結果、GPT-5.5 は Anthropic 社の「Claude Mythos Preview」に匹敵し、一部の指標ではそれを上回る高度なサイバー攻撃能力を持つことが明らかになりました[1, 4, 10]。

評価の概要と主な結果

AISI は、GPT-5.5 の能力を 2 つの主要な方法で評価しました[1, 7]。

Capture The Flag（CTF）形式の評価 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 CTF 形式のタスク評価では、GPT-5.5 は非常に高い性能を示しました[1, 7]。

- 95 種類のタスクを用いた評価において、最難関の「Expert」レベルで平均成功率 71.4%を記録しました。これは、以前に評価された Claude Mythos Preview の 68.6%を上回る数値です[1, 7, 8, 13]。
- 具体的な事例として、人間の専門家が専用ツールを使っても約 12 時間かかるとされるリバースエンジニアリング

の課題を、GPT-5.5 は人間の補助なしに約 11 分で完了しました。その際の API 利用コストはわずか 1.73 ドルでした[1, 7]。

マルチステップ攻撃シミュレーション 実際の企業ネットワーク環境を模した「Cyber Range」でのシミュレーションでは、GPT-5.5 は複数の手順を要する複雑な攻撃を自律的に実行する能力を示しました[1, 7]。

- 人間の専門家が手動で約 20 時間かかると推定される企業ネットワークへの侵入シミュレーションにおいて、GPT-5.5 は 10 回の試行のうち 2 回、完全に自律的な成功を収めました[7, 8]。
- これにより、GPT-5.5 は Claude Mythos Preview に続き、この種の高度なシミュレーションをエンドツーエンドで完了した 2 番目の AI モデルとなりました[1, 8, 9]。

AISI の見解と今後の課題

AISI は、今回の結果が特定のモデルに限った話ではなく、フロンティア AI 全体の能力向上のトレンドを反映している可能性が高いと指摘しています[1, 4, 8]。AI の推論能力、自律性、コーディング能力といった一般的な性能の向上が、副次的にサイバー攻撃能力をも高めているという見方です[1, 4]。

また、AISI は GPT-5.5 の安全対策（セーフガード）についても評価を行いました。専門家によるレッドチームテストの結果、GPT-5.5 のセーフガードを回避できる「ユニバーサル・ジェイルブレイク」が 6 時間で開発されたことも報告されています。OpenAI はその後、この問題に対処するための更新を行いました 8 。

AI Security Institute（AISI）について

AISI は、英国の科学・技術・イノベーション省（DSIT）内に設置された政府系研究機関で、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で唯一の公的機関です[4, 11]。AISI の評価は、統一された環境下で複数のモデルを比較できる政府公式のベンチマークとして、客観性と信頼性が高いとされています 4 。

1. GPT－5.5 のサイバー攻撃能力、一部で「Mythos」上回る
2. Read our full evaluation:
3. GPT－5.5、Claude Mythos Preview に続き企業ネットワーク ...
4. Mythos を上回った GPT-5.5 の実力、英国 AISI が明かした AI ...
5. GPT－5.5 ― 英国政府が初めて数字にした AI サイバー能力の現実
6. UK AISI Says GPT－5.5 Is One of the Strongest Cyber ...
7. UK AI Safety Institute warns GPT－5.5 cyber threat matches ...
8. Our evaluation of OpenAI's GPT－5.5 cyber capabilities
9. AISI Blog ｜ The AI Security Institute
10. GPT－5.5 の高度なサイバー攻撃能力に警鐘 「ミトス」に続く 2 例 ...

11. Frontier AI Trends Report by The AI Security Institute （AISI）
12. 英 AI 安全研究所、GPT－5.5 の高いサイバー攻撃能力を報告
13. GPT－5.5 のサイバー攻撃能力は一部「Mythos 超え」 英政府 ...
14. AI Security Institute – Frontier AI Trends report factsheet
15. UK AISI Releases Cyber Agent Evaluation Ranges
16. The AI Security Institute （AISI）
17. International AI Safety Report

GPT-5.5 のサイバーセキュリティ能力はどのように評価されましたか？

英国の AI Security Institute（AISI）によって、OpenAI の「GPT-5.5」のサイバーセキュリティ能力は、主に 2 つのアプローチで評価されました[7, 9, 15]。

1. Capture The Flag（CTF）形式の評価

これは、システムの脆弱性を発見し、隠された情報を奪取する能力や攻撃コードの実行能力を測定するためのタスク群です[7, 9, 12, 15]。

- **評価内容**: 難易度別に設定されたタスクで、GPT-5.5 の成功率が測定されました 9 。
- **主な結果**:
    - 最も難易度の高い「Expert」レベルのタスク群において、GPT-5.5 は平均 71.4%の成功率を記録しました[7, 9, 10, 23]。これは、以前に評価された Anthropic 社の「Claude Mythos Preview」の 68.6%を上回る数値です[7, 9, 23]。
    - 具体的な事例として、人間のセキュリティ専門家が約 12 時間かかるリバースエンジニアリングの課題を、GPT-5.5 はわずか 10 分 22 秒で解決したと報告されています[7, 8, 24]。

2. マルチステップ攻撃シミュレーション（サイバーレンジ）

これは、実際の企業ネットワーク環境と防御システムを模した、より実践的なテストです[7, 15]。複数の手順を要する複雑な攻撃を自律的に実行できるかを評価します。

- **評価内容**: 企業ネットワークへの侵入を模した 32 段階のシミュレーション「The Last Ones」などが用いられました[7, 8]。

- **主な結果**:
    - 人間の専門家が手動で約 20 時間かかるとされるこのシミュレーションにおいて、GPT-5.5 は 10 回の試行のうち 2 回、完全に自律的な攻略に成功しました[7, 8]。
    - これにより、GPT-5.5 は Claude Mythos Preview に続き、この種の高度なシミュレーションをエンドツーエンドで完了した 2 番目の AI モデルとなりました[14, 16]。

これらの評価から、AISI は GPT-5.5 が非常に高いサイバー攻撃能力を有していると結論付けました。また、異なる開発企業から同レベルの能力を持つモデルが相次いで登場していることから、これが AI 業界全体のトレンドである可能性を指摘しています[7, 9, 13, 17]。

1. Our evaluation of OpenAI's GPT－5.5 cyber capabilities
2. Read our full evaluation:
3. 「GPT－5.5」の検索結果 － Yahoo！ニュース
4. GPT－5.5 － 英国政府が初めて数字にした AI サイバー能力の現実
5. 「GPT 5.5」の検索結果 － Yahoo！ニュース
6. Our evaluation of OpenAI's GPT－5.5 cyber capabilities
7. GPT－5.5 の高度なサイバー攻撃能力に警鐘 「ミトス」に続く 2 例 ...
8. AISI レポート：OpenAI の GPT－5.5、Claude Mythos と同レベル ...
9. GPT－5.5 のサイバー攻撃能力は一部「Mythos 超え」 英政府 ...
10. 英 AI 安全研究所、GPT－5.5 の高いサイバー攻撃能力を報告
11. 【2026 年 4 月速報】GPT－5.5－Cyber 完全解説｜TAC 申請・活用法
12. 英政府、GPT－5.5 の高度なサイバー攻撃能力に警鐘 「ミトス」に ...
13. 研究動向 － ITmedia AI＋
14. Our evaluation of OpenAI's GPT－5.5 cyber capabilities
15. GPT－5.5 のサイバー攻撃能力、一部で「Mythos」上回る
16. AISI Blog ｜ The AI Security Institute
17. GPT－5.5・Claude Opus 4.7、推論能力は人間並みか…ARC ...
18. OpenAI の新モデル「GPT－5.4－Cyber」、サイバーセキュリティの ...
19. GPT－5.5 のサイバーテストで OpenAI は Anthropic Mythos に迫る
20. GPT－5.5 が「ネットワーク完全乗っ取り攻撃」を自律的に成功
21. Claude Mythos vs GPT－5.5: Enterprise AI Security ...
22. GPT－5.5 System Card － Deployment Safety Hub － OpenAI
23. AI Security Institute: GPT－5.5 "may be the strongest model ...
24. GPT－5.5 Solved a 12－Hour Reverse Engineering ...

25. GPT－5.4－Cyber and GPT－5.5 for security
26. AI セーフティ・インスティテュート（AISI）