AISI による GPT-5.5 サイバー評価の検証レポート

ChatGPT

Executive Summary

結論から言うと、ユーザー前提の中核部分――「英国政府系の AISI が 2026 年 4 月 30 日に OpenAI の GPT-5.5 のサイバーセキュリティ能力評価を公表した」という点――は、一次情報で**概ね確認できました**。確認できた公式一次資料は、AISI 公式サイト上のブログ記事 “Our evaluation of OpenAI's GPT-5.5 cyber capabilities” で、日付は Apr 30, 2026 と表示されています。一方で、AISI **名義の個別プレスリリース、GPT-5.5 専用の報告書** PDF**、技術補遺** PDF **は本調査時点では未確認**です。AISI は英国の Department for Science, Innovation and Technology[1] 内の研究組織として案内されています。[2]

AISI の公式評価によれば、GPT-5.5 は 95 件の狭義サイバー課題群と 2 種類のサイバーレンジで評価され、最難関の Expert 級狭義課題では**平均成功率** 71.4%±8.0%を記録しました。また、32 ステップの企業ネットワーク攻撃シミュレーションを 2/10 **回**でエンドツーエンド完遂したと AISI は後日補正しています。もっとも、OpenAI の公式 System Card には依然として 1/10 **回**と記載されており、一次ソース間で重要な数値差が残っています。評価環境は統制された研究設定であり、**アクティブ防御者、実運用の防御ツール、アラート・ペナルティは欠如**しているため、よく防御された現実システムへの一般化には限界があります。[3]

安全対策面では、AISI は GPT-5.5 のサイバー safeguards に対して**全ての悪性サイバー問い合わせで違反出力を引き出す** universal jailbreak を 6 時間の専門家レッドチーミングで作成できたと報告しました。OpenAI はその後 safeguard stack を複数回更新したと述べていますが、AISI は**最終構成の有効性を検証できなかった**としています。他方で OpenAI は、自社の別の外部レッドチーミングでは**最終ローンチ構成で検証済み高重大度のサイバージェイルブレイクは全てブロックされた**と主張しており、ここでも評価主体・構成差に由来する食い違いがあります。[4]

総合すると、本件は「GPT-5.5 が直ちに一般公開環境で自律サイバー攻撃を再現する」と読むべき報告ではありません。しかし、「高度な推論・自律性・コーディング能力の一般的向上が、そのまま攻撃側サイバー能力の底上げを生んでいる」というAISI の含意は重いです。英国政府・NCSC・OpenAI はいずれも、公開全面禁止ではなく、**段階的アクセス制御、監視、**Trusted Access**、パッチ運用強化**へ舵を切っています。[5]

主要発見

- AISI **公式発表の存在は確認できた**。ただし確認できたのは AISI 公式ブログ記事であり、AISI 名義の別個のプレスリリースや GPT-5.5 専用 PDF 報告書は本調査範囲では未確認です。[6]
- AISI 公式ブログでの主要数値は、Expert 級狭義課題で 71.4%±8.0%、32 ステップ企業ネットワーク攻撃レンジの 2/10 **完遂**、`rust_vm` 課題の 10 **分** 22 **秒・**$1.73 です。[7]
- OpenAI **公式** System Card **との不一致**がある。AISI は企業ネットワークレンジの完遂回数を 2/10 へ修正したが、OpenAI の System Card は 1/10 のままです。[8]
- **評価対象の表現にも差異**がある。AISI は「GPT-5.5 の early checkpoint」と記述し、OpenAI は「representative launch checkpoint」および「reduced refusals checkpoint」を UKAISI へ提供したと記述しています。モデル ID やハッシュは未公開です。[9]
- **再現性は部分的**です。AISI は最小限の ReAct エージェント、Kali Linux、Inspect AI、context compaction などを公開し、方法論説明は一定程度ありますが、狭義課題群やサイバーレンジの完全アーティファクトは公開されていません。[10]
- **安全対策の評価結果は深刻**です。AISI は universal jailbreak を 6 時間で発見し、最終構成の再検証はできなかった一方、OpenAI は別の外部レッドチームで最終構成の高重大度 jailbreak を遮断できたと説明しています。[11]

- 英国政府側は本件を単体ニュースとしてよりも、Breaches Survey、**企業向け公開書簡**、Cyber Security and Resilience Bill、NCSC の“vulnerability patch wave”**準備**と結び付けて位置付けています。[12]
- 国際的には、**高度サイバーAI のアクセス管理そのものが政策論点化**しています。OpenAI は Trusted Access を拡大し、報道では GPT-5.5-Cyber を「critical cyber defenders」へ限定展開するとされ、欧州では他社高性能モデルへの公的アクセス確保を求める動きも見られます。[13]

一次情報の確認

本件の一次情報として最も重要なのは、AISI 公式ブログ記事です。文書名は “Our evaluation of OpenAI's GPT-5.5 cyber capabilities” で、ページ上の表示日付は Apr 30, 2026 です。AISI ブログ一覧にも同記事が掲載されており、AISI 研究ページ・既存論文へのリンクは確認できる一方、GPT-5.5 **専用の** AISI PDF **報告書や技術補遺** PDF **への公式リンクは確認できませんでした**。OpenAI 側の一次資料としては GPT-5.5 System Card の Web 版と PDF 版が確認でき、公開日は April 23, 2026、更新日は April 24, 2026 です。AISI 評価そのものを主題にした別個の gov.uk プレスリリースは、本調査範囲では未確認でした。[14]

文書種別	入手可否	文書名	発表・更新日時	URL	PDF等	判定
AISI 公式記事	可能	Our evaluation of OpenAI's GPT-5.5 cyber capabilities	2026-04-30（時刻は未確認）	`https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities`	PDF未確認	**確認済み**
AISI	可	AISI Blog listing	2026-04-	`https://www.aisi.gov.uk/blog`	PDF	**確**

文書種別	入手可否	文書名	発表・更新日時	URL	PDF等	判定
公式一覧掲載	可能		30掲載確認		不要	**認済み**
AISIのGPT-5.5専用技術補遺	未確認	該当文書を確認できず	未確認	未確認	未確認	**未確認**
AISIの方法論共有研究ページ	可能	Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios	2026-03-16	https://www.aisi.gov.uk/research/measuring-ai-agents-progress-on-multi-step-cyber-attack-scenarios	arXiv PDFあり	**確認済み**
OpenAI公式System Card	可能	GPT-5.5 System Card	公開2026-04-23、更新2026-04-24	https://openai.com/index/gpt-5-5-system-card/	PDFあり	**確認済み**
OpenAI Syste	可能	GPT-5.5 System Card PDF	2026-04-23	https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf	PDF	**確認**

文書種別	入手可否	文書名	発表・更新日時	URL	PDF等	判定
m Card PDF						**済み**
gov.uk 個別プレスリリース	未確認	GPT-5.5 評価単独を主題とする文書	未確認	未確認	未確認	**未確認**

上表のうち「未確認」は、存在しないと断定する意味ではなく、2026-05-03 **時点で公開一次情報として確認できなかった**ことを意味します。特に AISI 記事には日付表示はあるものの、**時刻表示は見当たらず、正確な発表時刻は未確認**です。[15]

評価内容の分解

AISI がこの評価で問おうとしている中心命題は、「4 月上旬に Anthropic[16] の Claude Mythos Preview で観測された高い攻撃的サイバー能力が、そのモデル固有の飛躍なのか、それともより広い能力トレンドなのか」という点です。AISI は GPT-5.5 の結果について、“a second model, from a different developer, now reaches a similar level of performance” と整理しており、個別モデルの特異点というより、汎用能力の進展に伴う横断的トレンドとして読んでいます。[17]

評価対象モデルの記述には注意が必要です。AISI ブログは GPT-5.5 を“an early checkpoint” と表現する一方、OpenAI の System Card は UKAISI に対して“a representative launch checkpoint and a checkpoint with reduced refusals” を提供したと記しています。したがって、「一般公開時の最終商用設定そのもの」をそのまま試験したと断定するのは避けるべきです。**バージョン識別子、重みハッシュ、**

API revision、safeguard **構成差分**は公開されておらず、ここは再現性判断のボトルネックです。[9]

方法論は二層構造です。第一に、AISI は 95 **件の狭義サイバー課題**を 4 段階難度で保有し、そのうち高度課題は 27 **件の** Practitioner と 21 **件の** Expert から成ります。高度課題は Crystal Peak Security[18] と Irregular[19] と協力して構築され、脆弱性調査、エクスプロイト開発、暗号解析、マルウェア解析などを問います。AISI ブログの 71.4%はこの高度課題の**平均成功率**であり、OpenAI System Card が示す pass@1 66.7% や pass@5 90.5% とは**指標定義が異なる**ため、数値を横並び比較するときは注意が必要です。[20]

第二に、AISI は 2 種類のサイバーレンジを用います。企業ネットワーク攻撃レンジ “The Last Ones” は 32 ステップ、ICS レンジ “Cooling Tower” は 7 ステップです。これらのレンジは、既存論文によれば SpecterOps[21] と Hack The Box[22] が設計・構築に関与し、AISI は最小限の ReAct エージェント、Kali Linux、Bash・Python・Mythic コマンド、context compaction、Inspect AI フレームワークで実行したと説明しています。モデル性能そのものを見たいという意図から、scaffold は意図的に最小限に抑えられています。[23]

実験環境の限界は AISI 自身がかなり率直に開示しています。レンジには**アクティブ防御者がいない、検知トリガーに点数上の不利益がない、実環境より脆弱性密度が高い可能性がある、現実の企業ネットワークよりアーティファクト密度が低い**、そして評価は**既にネットワークアクセスを得た脆弱な標的に特定目的で向けたときに何ができるか**にスコープされている、という条件です。ゆえに、AISI の数値は「現実世界のうち、脆弱・弱防御環境に関する上限寄りの能力評価」であって、「強固な本番システムに対する成功確率」ではありません。[24]

再現性については、**部分的にあるが完全ではない**というのが妥当です。プラス材料として、AISI は Inspect AI、ReAct、context compaction、攻撃連鎖の高レベル記述を公開し、Appendix C では攻撃チェーンの概要も示しています。マイナス材料として、狭義課題 95 件の完全セット、サイバーレンジの完全イメージ、モデルチェックポイント識別子、実行ログ、最終 safeguard 構成は公開されていません。論文自体も

、再現性と攻撃手引き化のトレードオフを明示しています。したがって、**第三者がAISI の数値を厳密再現する見込みは現時点では限定的**です。[25]

主な結果と比較表

AISI 一次資料から確実に引ける定量結果は、主として次の 5 群です。狭義高度課題では GPT-5.5 が Expert 級で 71.4%±8.0%を記録し、AISI は「試験した中で最強の可能性がある」と評価しました。企業ネットワークレンジでは AISI が 2/10 **完遂**へ訂正、OpenAI は 1/10 のまま、ICS レンジは未完遂です。個別課題 `rust_vm` では人間専門家が約 12 時間かかる難題を 10 分 22 秒・$1.73 で突破しました。safeguard 面では universal jailbreak が全悪性クエリで通った一方、OpenAI の内製・別外部テストでは最終構成で高重大度 cyber jailbreak を遮断できたとされます。AISI 公式資料には、**誤検知率・偽陽性率**・full confusion matrix・TLO **全体の金額コスト**は掲載されていません。[26]

評価対象	攻撃手法・シナリオ	主な定量結果	影響度の読み	緩和策・運用上の示唆	出典
狭義高度課題 Expert	CTF 形式の脆弱性調査・エクスプロイト・暗号・RE	71.4%±8.0%	Frontier モデル群の中で最上位圏。AISI は「最強の可能性」と評価	公開ベンチだけでなく長い予算・複数反復が必要	AISI 公式 [27]

評価対象	攻撃手法・シナリオ	主な定量結果	影響度の読み	緩和策・運用上の示唆	出典
OpenAI 記載の UKAISI Expert 指標	同上だが指標定義が異なる	pass@5 = 90.5% ± 12.9%、pass@1 = 66.7% ± 15.9%	AISI ブログの 71.4%とは別指標。直接比較は不可	指標差を明記して読む必要	OpenAI 公式 [28]
低難度課題群	lower-difficulty/basic tasks	lower-difficulty で 100%、basic は少なくとも 2026 年 2 月以降ほぼ飽和	初級〜中級サイバー技能は多くの frontier モデルで頭打ち	より高難度・長時間タスクへの評価重点化が必要	OpenAI 公式/AISI 公式 [29]
企業ネットワークレンジ “The Last Ones”	32-step 企業ネットワーク攻撃	AISI 更新値 2/10 完遂。OpenAI System Card は 1/10 のまま	弱防御・既侵入前提なら end-to-end 攻撃能力が出始めている可能性	強固環境への一般化は不可。監	AISI 公式/OpenAI 公式 [8]

評価対象	攻撃手法・シナリオ	主な定量結果	影響度の読み	緩和策・運用上の示唆	出典
				視・分離・検知前提で読むべき	
ICS レンジ “Cooling Tower”	7-step 産業制御系攻撃	完遂できず。AISI は「どのモデルも未完遂」と記述	ICS 攻撃能力を高く見積もるのは時期尚早	OT 防御の特殊性はなお有効。IT 側侵入面の防御が重要	AISI 公式/先行論文 [30]
`rust_vm` 個別課題	カスタム VM の RE、逆アセンブル、制約解決	人間約 12 時間、GPT-5.5 は 10 分 22 秒、$1.73	個別技能ではすでに人間専門家の大幅短縮が起きている	RE・マルウェア解析・	AISI 公式 [31]

評価対象	攻撃手法・シナリオ	主な定量結果	影響度の読み	緩和策・運用上の示唆	出典
				脆弱性調査の AI 悪用/防御利用双方に注意	
cyber safeguards	malicious cyber queries への安全対策	AISI: universal jailbreak が全クエリで成立、開発 6 時間	safeguard 破りの汎用化リスクが現実	多層防御、構成管理、継続的レッドチームが必須	AISI 公式/OpenAI 公式 [4]
OpenAI の最終構成主張	外部レッドチームの高重大度 jailbreak	final launch configuration では verified high-severity cyber jailbreaks を全て	AISI の未検証問題とは別系統の主張	実運用では構成差	OpenAI 公式 [32]

評価対象	攻撃手法・シナリオ	主な定量結果	影響度の読み	緩和策・運用上の示唆	出典
		遮断		分の監査が鍵	

未公表・未確認の指標も明確です。AISI 公式評価には、偽陽性率、誤検知率、シナリオ別 misuse probability、攻撃オペレーション全体の期待費用は出ていません。OpenAI は自社のサイバー安全学習評価として、Production data で 0.928、Synthetic data で 0.975 の policy compliance rate を示していますが、これは AISI 評価の誤検知率ではなく、OpenAI **独自の** safety eval です。[33]

なお、レンジの「人間専門家が要する時間」には一次ソース間で揺れがあります。AISI の 3 月論文・研究ページでは The Last Ones **＝約 14 時間**と見積もる一方、GPT-5.5 ブログと OpenAI System Card では **約 20 時間**とされています。推計更新・レンジ改訂・記述差のいずれかは**未確認**であり、厳密な比較の際はこの点を脚注扱いにすべきです。[34]

AISI の結論と OpenAI・第三者の反応

AISI の公式結論は、単に「GPT-5.5 が強い」ではありません。より重要なのは、**攻撃側サイバー能力の急速な改善が、個別モデル専用の学習ではなく、長期自律・推論・コーディング能力の一般的進歩の副産物かもしれない**という点です。だから AISI は、近い将来に**さらに短い間隔で能力上昇が起こりうる**と示唆しています。同時に AISI は、結果を過剰解釈しないよう、現行レンジにはアクティブ防御者や防御ツールがなく、well-defended target **への成功はこの結果からは断言できない**と繰り返し留保しています。[24]

AISI の推奨は、公開停止一辺倒ではありません。AISI は、より強力なモデルが Trusted Access 経由を含め広く利用可能になるなら、**防御側も同じ能力を自組織の保護に使うべき**だと述べています。この文脈で AISI は、英国の National Cyber Security Centre[35] と連携した防御側活用のブログ、Breaches Survey、企業向け公開書簡、サイバー・レジリエンス法案、そして“vulnerability patch wave” への備えを参照しています。政策的には「閉じる」より「段階的に開くが、防御側に優先配分し、運用上の基本衛生を強める」という方向です。[36]

OpenAI 側の公式コメントは三つに整理できます。第一に、GPT-5.5 のサイバー能力は Preparedness Framework 上で High **だが** Critical **ではない**と位置付けています。第二に、モデルの一般公開に際しては**最強の** safeguards、会話モニタ、アカウント単位執行、Trusted Access for Cyber、より許容的な cyber-permissive access を組み合わせるとしています。第三に、AISI が見つけた universal jailbreak については修正を進めたうえで、別系統の外部レッドチームでは**最終ローンチ構成で高重大度** cyber jailbreak **を全件遮断できた**と主張しています。[37]

ただし、この OpenAI 説明にはなお未解消点があります。AISI は**最終構成を検証できなかった**とし、OpenAI の System Card は 2026-04-24 更新以降も 1/10 **完遂**表記のままです。つまり OpenAI は AISI の方向性を否定してはいないものの、**修正済み数字の反映・構成検証の透明性・対象** checkpoint **の一意性**では十分とは言いがたい状態です。反論というより、「評価は受け入れるが、最終運用では追加保護がある」とする統治的応答に近いです。[38]

第三者反応としては、主要メディアは総じて「Mythos だけの特異点ではなく、frontier model 全体のトレンドだ」という AISI 解釈を強調しました。Ars Technica[39] は “not a breakthrough specific to one model” という論点を拾い、The Decoder[40] は、GPT-5.5 が Mythos に並ぶレベルの攻撃能力を示したこと、かつ防御のない環境での結果に留意すべきことを強調しています。The Verge[41] は、OpenAI が GPT-5.5-Cyber を“critical cyber defenders”に限定配布すると報じ、アクセス統制が新しい業界標準になりつつあることを示しました。Reuters[42] は、同時期に欧州側が他社の

高性能サイバーAI へのアクセスを求める動きを報じており、**モデル評価そのものが安全保障上のアクセス問題に接続している**ことを示しています。[43]

独立再現については、本調査範囲では未確認です。確認できた第三者公開物の大半は報道・解説であり、AISI や OpenAI 以外の学術機関・セキュリティ企業が、GPT-5.5 について同一レンジや同等条件で再試験した公開一次資料は見当たりませんでした。少なくとも AISI 公式、OpenAI 公式、arXiv 上の直接資料、主要セキュリティ企業公開ページの範囲では、**公開再現実験は未確認**と評価するのが妥当です。[44]

信頼性評価と日本・国際的含意

信頼性評価の**強み**は明確です。第一に、AISI 公式ブログ、OpenAI 公式 System Card、AISI の先行方法論論文、gov.uk の政策文書、NCSC の補完ガイダンスという**一次ソース連鎖が揃っている**点です。第二に、AISI は限界を隠さず、アクティブ防御不在、alert penalty 不在、脆弱性密度、ネットワークアクセス前提などを列挙しています。第三に、方法論の継続性があり、3 月のレンジ論文・5 月の評価で同系列の instrument を使って能力推移を追っている点です。第四に、平均値だけでなく標準誤差や予算条件を示しており、最低限の数量的手がかりがあります。[45]

一方で**弱み・バイアス**もあります。最大の弱みは、checkpoint **の一意性がない**こと、**レンジ・課題群の完全公開がない**こと、**現実ネットワークに比べて簡略化された環境**であることです。また、AISI は開発企業から事前アクセスを受けて評価しており、完全独立のブラインド監査ではありません。もっとも AISI 先行論文は、レンジの設計・実行・解釈責任は著者側にあると明記しています。さらに、本件では 1/10 **対** 2/10、14 **時間対** 20 **時間**という公的文書間のズレがあり、ヘッドライン数字の厳密性は少し落ちます。ただし、これらのズレは「能力が上がっている」という大勢を覆すほどではありません。[46]

日本向け実務含意は、公式な対 GPT-5.5 利用制限の有無よりも、**運用設計をどう変えるか**にあります。英国の公開書簡と NCSC の“vulnerability patch wave”論は、AI が脆弱性探索・攻撃連鎖・修正提案を高速化する世界では、従来の「月例パッチ」「属人的トリアージ」「脆弱なログ設計」では追いつかなくなるという前提に立って

います。日本企業にとっての実務的示唆は、少なくとも以下の四点です。**第一に、パッチ適用能力そのものを経営課題に昇格すること。第二に、セキュリティログと検知基盤を、AI 生成ノイズも含めて運用できるよう整備すること。第三に、社内でサイバー用途に LLM を使う場合は、一般公開設定と permissive access を明確に分け、承認・監査・保存方針を設けること。第四に、RE・脆弱性調査・マルウェア解析を自動化する防御ユースを、限定環境・権限分離・契約管理の下で試験導入すること**です。これは英国政府文書を日本向けに敷衍した推論ですが、かなり実務的です。[47]

国際的には、**アクセス統制が新たなスタンダードになりつつある**とみてよいでしょう。OpenAI は Trusted Access for Cyber を拡大し、報道では GPT-5.5-Cyber の限定提供が示されました。Reuters が報じたように、欧州では高性能サイバーAI への公的アクセスそのものを求める議論が進んでいます。これは今後、単なるモデル性能の話から、**誰が、どの条件で、どのモデルにアクセスできるのか**という統治・競争・安全保障の論点にシフトすることを意味します。[48]

以下のタイムラインは、本件評価を、英国政府・OpenAI・サイバー政策の流れの中に置き直したものです。時刻が取れない項目は日付のみを示しています。タイムラインの各出来事の事実関係は、直後の参考資料一覧に挙げる一次ソースに基づきます。[49]

```
timeline
    title GPT-5.5 サイバー評価をめぐる主要動向
    2026-04-14 : OpenAI が Trusted Access for Cyber 拡大と GPT-5.4-Cyber を公表
    2026-04-15 : 英政府高官が企業向け「AI cyber threats」公開書簡
    2026-04-23 : OpenAI が GPT-5.5 と System Card を公開
    2026-04-24 : OpenAI が System Card を API safeguards 情報で更新
    2026-04-30 : AISI が GPT-5.5 cyber capabilities 評価を公表
    2026-05-01 : NCSC が“vulnerability patch wave”への備えを呼びかけ
```

Open questions and limitations

- AISI が GPT-5.5 専用の**技術補遺** PDF や完全報告書を別途公開しているかは、**未確認**です。[6]
- AISI 公式記事の**時刻**は取れておらず、発表日時は 2026-04-30（**日付のみ確認**）です。[50]
- AISI の 2/10 修正値が、OpenAI の System Card や PDF へいつ反映されるかは**未確認**です。[8]
- AISI が評価した checkpoint と、一般公開・API・Trusted Access の各構成の**厳密な差分**は未公開です。[51]
- 本件について、AISI/OpenAI 以外の第三者による**公開再現実験**は本調査範囲では未確認です。[52]

参考資料一覧

優先度は A＝**一次ソース**, B＝**準一次ソース・方法論**, C＝**報道・解説** としました。

- A-1 AISI 公式記事「Our evaluation of OpenAI's GPT-5.5 cyber capabilities」
  URL: `https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities` [17]

- A-2 OpenAI 公式「GPT-5.5 System Card」
  URL: `https://openai.com/index/gpt-5-5-system-card/` [53]

- A-3 OpenAI Deployment Safety Hub PDF「GPT-5.5 System Card」
  URL: `https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf` [54]

- A-4 OpenAI 公式「Introducing GPT-5.5」
  URL: `https://openai.com/index/introducing-gpt-5-5/` [55]

- A-5 OpenAI 公式「Trusted access for the next era of cyber defense」
  URL: `https://openai.com/index/scaling-trusted-access-for-cyber-defense/` [56]

- A-6 gov.uk「AI cyber threats: open letter to business leaders」
  URL: `https://www.gov.uk/government/publications/ai-cyber-threats-`

open-letter-to-business-leaders/ai-cyber-threats-open-letter-to-business-leaders-html [57]

- A-7 gov.uk「Cyber security breaches survey 2025/2026」
  URL: https://www.gov.uk/government/statistics/cyber-security-breaches-survey-20252026/cyber-security-breaches-survey-20252026 [58]
- A-8 gov.uk「Call to action for AI companies to work with UK Government on national cyber defence」
  URL: https://www.gov.uk/government/news/call-to-action-for-ai-companies-to-work-with-uk-government-on-national-cyber-defence [59]
- A-9 NCSC「Preparing for a 'vulnerability patch wave'」
  URL: https://www.ncsc.gov.uk/blogs/prepare-for-vulnerability-patch-wave [60]
- A-10 NCSC「Why cyber defenders need to be ready for frontier AI」
  URL: https://www.ncsc.gov.uk/blogs/why-cyber-defenders-need-to-be-ready-for-frontier-ai [61]
- B-1 AISI 研究ページ「Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios」
  URL: https://www.aisi.gov.uk/research/measuring-ai-agents-progress-on-multi-step-cyber-attack-scenarios [62]
- B-2 arXiv「Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios」
  URL: https://arxiv.org/abs/2603.11214 / https://arxiv.org/pdf/2603.11214 [63]
- B-3 AISI 公式「Evidence for inference scaling in AI cyber tasks」
  URL: https://www.aisi.gov.uk/blog/evidence-for-inference-scaling-in-ai-cyber-tasks-increased-evaluation-budgets-reveal-higher-success-rates [64]
- C-1 Reuters「EU should seek access to Anthropic's Mythos, Bundesbank says」
  URL: https://www.reuters.com/legal/litigation/eu-should-seek-access-anthropics-mythos-bundesbank-says-2026-04-29/ [65]

- C-2 The Verge「OpenAI's new security model is for 'critical cyber defenders' only」
  URL: `https://www.theverge.com/ai-artificial-intelligence/921073/openai-sam-altman-new-cybersecurity-model-gpt-5-5-cyber` [66]
- C-3 Ars Technica「GPT-5.5 matches heavily hyped Mythos Preview in new cybersecurity tests」
  URL: `https://arstechnica.com/ai/2026/05/amid-mythos-hyped-cybersecurity-prowess-researchers-find-gpt-5-5-is-just-as-good/` [67]
- C-4 The Decoder「GPT-5.5 matches Claude Mythos in cyber attack tests, UK AI Security Institute finds」
  URL: `https://the-decoder.com/gpt-5-5-matches-claude-mythos-in-cyber-attack-tests-uk-ai-security-institute-finds/` [68]

☒ navlist ☒関連ニュース☒ turn40news32,turn40news31 ☒

---

[1] [2] [3] [4] [5] [6] [7] [8] [9] [12] [14] [15] [17] [18] [20] [21] [22] [24] [26] [27] [30] [31] [36] [39] [40] [41] [42] [44] [50] https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities

https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities

[10] [23] [25] [45] https://arxiv.org/html/2603.11214v1

https://arxiv.org/html/2603.11214v1

[11] [32] [33] [38] [54] https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf

https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf

[13] [55] https://openai.com/index/introducing-gpt-5-5/

https://openai.com/index/introducing-gpt-5-5/

[16] [65] https://www.reuters.com/legal/litigation/eu-should-seek-access-anthropics-mythos-bundesbank-says-2026-04-29/

https://www.reuters.com/legal/litigation/eu-should-seek-access-anthropics-mythos-bundesbank-says-2026-04-29/

[19] [53] https://openai.com/index/gpt-5-5-system-card/

https://openai.com/index/gpt-5-5-system-card/

[28] [29] [37] [46] [51] https://deploymentsafety.openai.com/gpt-5-5

https://deploymentsafety.openai.com/gpt-5-5

[34] https://www.aisi.gov.uk/blog/how-do-frontier-ai-agents-perform-in-multi-step-cyber-attack-scenarios

https://www.aisi.gov.uk/blog/how-do-frontier-ai-agents-perform-in-multi-step-cyber-attack-scenarios

[35] [58] https://www.gov.uk/government/statistics/cyber-security-breaches-survey-20252026/cyber-security-breaches-survey-20252026

https://www.gov.uk/government/statistics/cyber-security-breaches-survey-20252026/cyber-security-breaches-survey-20252026

[43] [67] https://arstechnica.com/ai/2026/05/amid-mythos-hyped-cybersecurity-prowess-researchers-find-gpt-5-5-is-just-as-good/

https://arstechnica.com/ai/2026/05/amid-mythos-hyped-cybersecurity-prowess-researchers-find-gpt-5-5-is-just-as-good/

[47] [60] https://www.ncsc.gov.uk/blogs/prepare-for-vulnerability-patch-wave

https://www.ncsc.gov.uk/blogs/prepare-for-vulnerability-patch-wave

[48] [49] [56] https://openai.com/index/scaling-trusted-access-for-cyber-defense/

https://openai.com/index/scaling-trusted-access-for-cyber-defense/

[52] [68] https://the-decoder.com/gpt-5-5-matches-claude-mythos-in-cyber-attack-tests-uk-ai-security-institute-finds/

https://the-decoder.com/gpt-5-5-matches-claude-mythos-in-cyber-attack-tests-uk-ai-security-institute-finds/

[57] https://www.gov.uk/government/publications/ai-cyber-threats-open-letter-to-business-leaders/ai-cyber-threats-open-letter-to-business-leaders-html

https://www.gov.uk/government/publications/ai-cyber-threats-open-letter-to-business-leaders/ai-cyber-threats-open-letter-to-business-leaders-html

[59] https://www.gov.uk/government/news/call-to-action-for-ai-companies-to-work-with-uk-government-on-national-cyber-defence

https://www.gov.uk/government/news/call-to-action-for-ai-companies-to-work-with-uk-government-on-national-cyber-defence

[61] https://www.ncsc.gov.uk/blogs/why-cyber-defenders-need-to-be-ready-for-frontier-ai

https://www.ncsc.gov.uk/blogs/why-cyber-defenders-need-to-be-ready-for-frontier-ai

[62] https://www.aisi.gov.uk/research/measuring-ai-agents-progress-on-multi-step-cyber-attack-scenarios

https://www.aisi.gov.uk/research/measuring-ai-agents-progress-on-multi-step-cyber-attack-scenarios

[63] https://arxiv.org/abs/2603.11214

https://arxiv.org/abs/2603.11214

[64] https://www.aisi.gov.uk/blog/evidence-for-inference-scaling-in-ai-cyber-tasks-increased-evaluation-budgets-reveal-higher-success-rates

https://www.aisi.gov.uk/blog/evidence-for-inference-scaling-in-ai-cyber-tasks-increased-evaluation-budgets-reveal-higher-success-rates

[66] https://www.theverge.com/ai-artificial-intelligence/921073/openai-sam-altman-new-cybersecurity-model-gpt-5-5-cyber

https://www.theverge.com/ai-artificial-intelligence/921073/openai-sam-altman-new-cybersecurity-model-gpt-5-5-cyber