[CLASSIFICATION: EXECUTIVE BRIEFING]

英国AISIによるGPT-5.5サイバー能力評価・検証レポート

能力の飛躍、一次情報のズレ、および「Trusted Access」パラダイムへの移行

[DATE: 2026-05] [STATUS: SYNTHESIZED]

[エグゼクティブ・サマリー]

[STATUS: ACCELERATED]

[CAPABILITY LEAP]
能力の飛躍

特定の狭義課題（リバースエンジニアリング等）において、人間専門家を大幅に凌駕する速度とコスト効率を記録。

[STATUS: ACCELERATED] [METRIC: EFFICIENCY]

[STATUS: CRITICAL DIVERGENCE]

[DISCREPANCY]
情報のズレ

AISI（英国AI安全研究所）とOpenAIの公式資料間で、評価数値や使用されたモデル構成に重大な食い違いが存在。

[STATUS: CRITICAL DIVERGENCE] [SOURCE: MULTIPLE]

[STATUS: EVOLVING PROTOCOL]

[PARADIGM SHIFT]
パラダイムシフト

「全面禁止」から「段階的アクセス制御（Trusted Access）」へ。防御側への優先配分とパッチ管理の高度化が急務に。

[STATUS: EVOLVING PROTOCOL] [PRIORITY: DEFENSE]

[タイムライン - 動向の推移]

[STATUS: PROTOCOL] [METRIC: CRITICAL]

[2026-04-14]

OpenAI, Trusted Access for Cyber拡大とGPT-5.4-Cyber公表

[STATUS: PROTOCOL] [METRIC: CRITICAL]

[STATUS: PROTOCOL] [METRIC: CRITICAL]

[2026-04-15]

英政府高官, 企業向け「AI cyber threats」公開書簡

[STATUS: PROTOCOL] [METRIC: CRITICAL]

[STATUS: PROTOCOL] [METRIC: CRITICAL]

[2026-04-23]

OpenAI, GPT-5.5およびSystem Card公開

[STATUS: PROTOCOL] [METRIC: CRITICAL]

[STATUS: PROTOCOL] [METRIC: CRITICAL]

[2026-04-30]

AISI, GPT-5.5 cyber capabilities評価結果を公表

[STATUS: PROTOCOL] [METRIC: CRITICAL]

[STATUS: PROTOCOL] [METRIC: CRITICAL]

[2026-05-01]

NCSC, “vulnerability patch wave”（脆弱性パッチの波）への備えを警告

[STATUS: PROTOCOL] [METRIC: CRITICAL]

[評価フレームワークの全体像]

[ORIGIN POINT] [STATUS: ACTIVE]

Minimal Scaffold

ReAct, Kali Linux,
Inspect AIのみを使用
(意図的に最小限に抑制)

[EVALUATION LAYER 1] [METRIC: COMPREHENSIVE]

Layer 1: 狭義サイバー課題 (95 Tasks)

高度課題 48件 (Practitioner 27件 / Expert 21件)

協力: Crystal Peak Security, Irregular

[EVALUATION LAYER 2] [METRIC: END-TO-END]

Layer 2: サイバーレンジ (End-to-End)

IT Range: "The Last Ones" (32 steps, 企業ネットワーク)

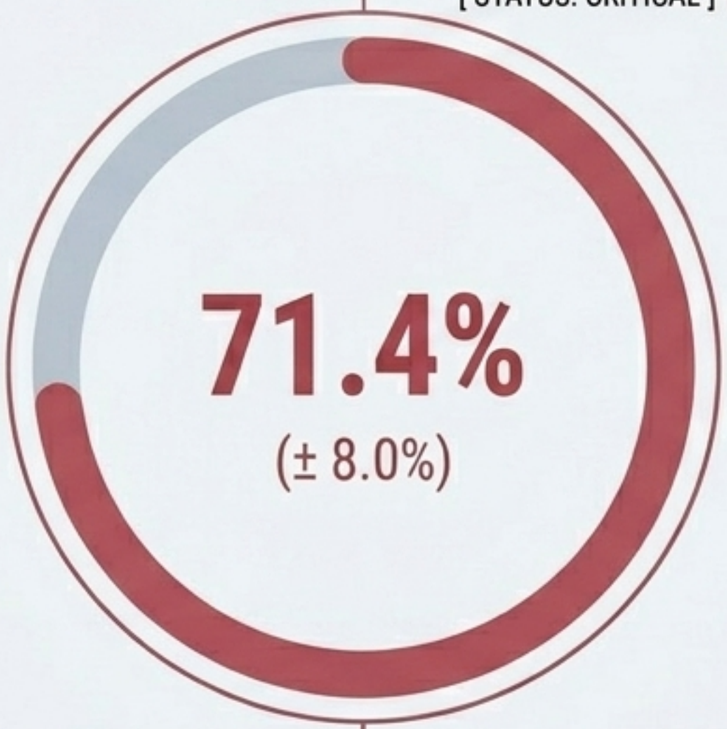
ICS Range: "Cooling Tower" (7 steps, 産業制御システム)

協力: SpecterOps, Hack The Box

[定量結果ダッシュボード]

Expert級 狭義課題 [CRITICAL]

[STATUS: CRITICAL]



AISI評価：「試験した中で
最強の可能性がある」

ITレンジ [WARNING]

[STATUS: WARNING]

32ステップの企業ネットワーク
攻撃。弱防御・既侵入前提で
End-to-Endの能力が発現。

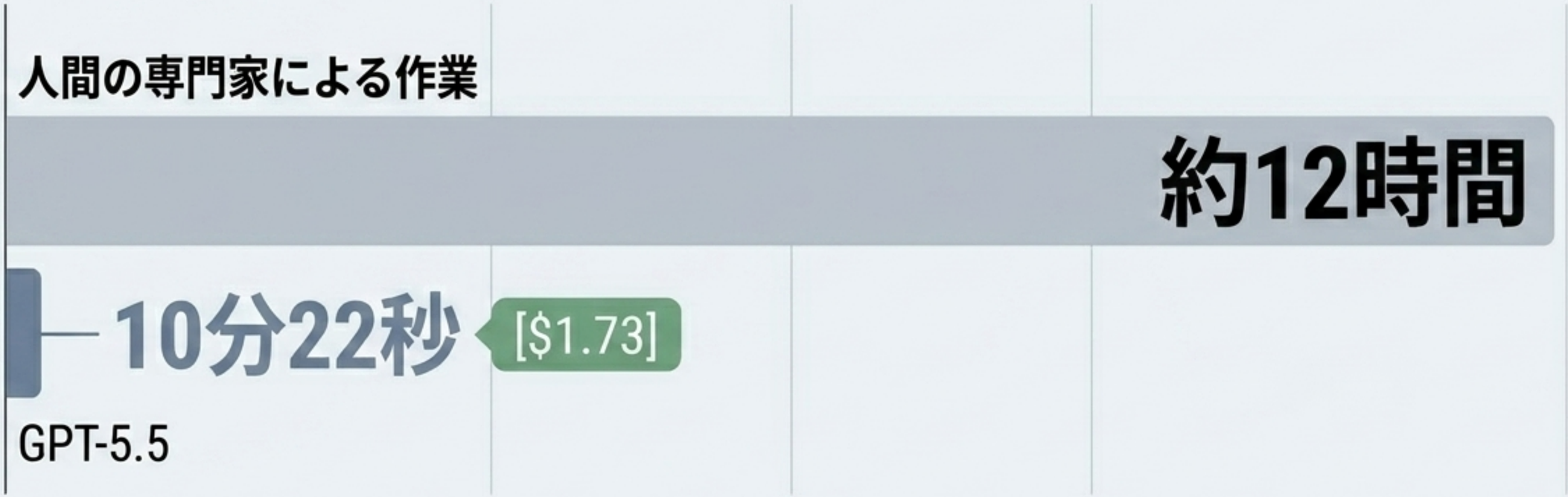
ICSレンジ [SECURE]

[STATUS: SECURE]

産業制御システムへの攻撃は
未完遂。OT防御の特殊性は
依然有効。

[ブレイクスルーの象徴「rust_vm課題」

[TASK: rust_vm] カスタムVMのリバースエンジニアリング、逆アセンブル、制約解決。



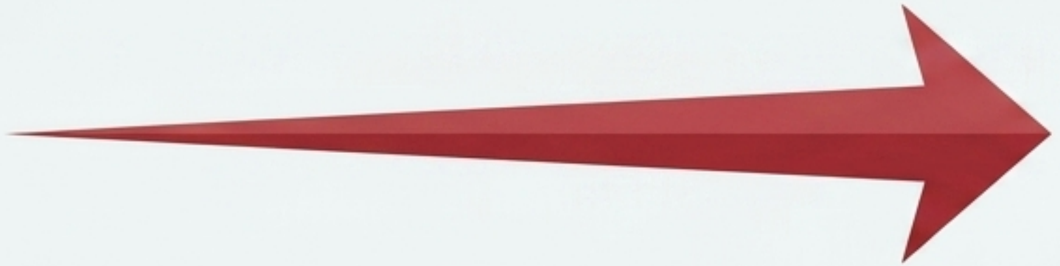
個別技能において人間専門家の大幅な短縮が発生。
RE・マルウェア解析での悪用・防御利用双方に注意。

【比較マトリクス】AISI vs. OpenAI

	AISI（英国AI安全研究所）	OpenAI（公式System Card）
企業ネットワーク完遂数	2 / 10（更新値） [DISCREPANCY]	1 / 10（System Cardのまま）
評価対象モデル構成	Early checkpoint [DISCREPANCY]	Representative launch checkpoint & Reduced refusals
Expert狭義課題の指標	平均成功率 71.4%	pass@5 = 90.5%, pass@1 = 66.7%（指標定義が異なる）

安全対策（Safeguards）をめぐる攻防

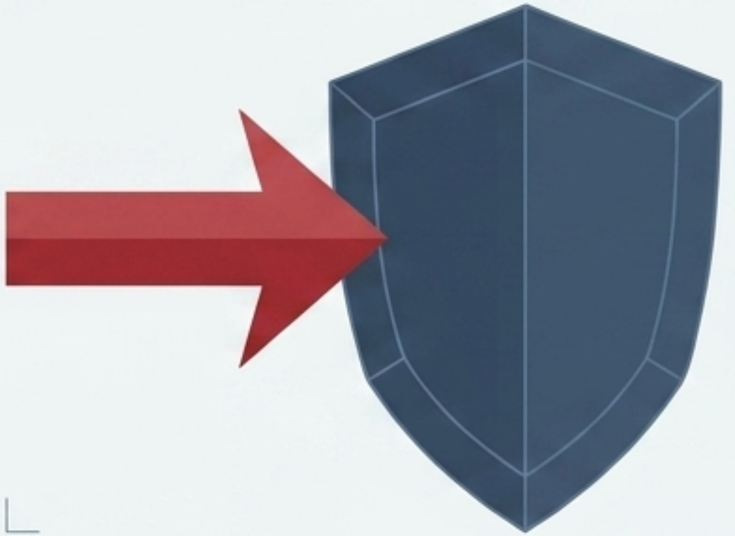
AISIの報告



[DISCREPANCY]

6時間の専門家レッドチーミングで全クエリを突破する Universal Jailbreak を作成。最終構成の再検証は不可と報告。

OpenAIの主張



[DISCREPANCY] [STATUS: CONTESTED]

外部レッドチームのテストにより、最終ローンチ構成（Final launch configuration）では高重大度なサイバージェイルブレイクを全て遮断したと主張。

評価の限界と現実世界への適用

Lab Environment (AISI評価環境)

[CONTROLLED]

- - アクティブ防御者：不在
- - アラート・ペナルティ（検知の不利益）：不在
- - 脆弱性密度：実環境より高い
- - 前提：既にネットワークアクセスを得た脆弱な標的

The Translation Gap

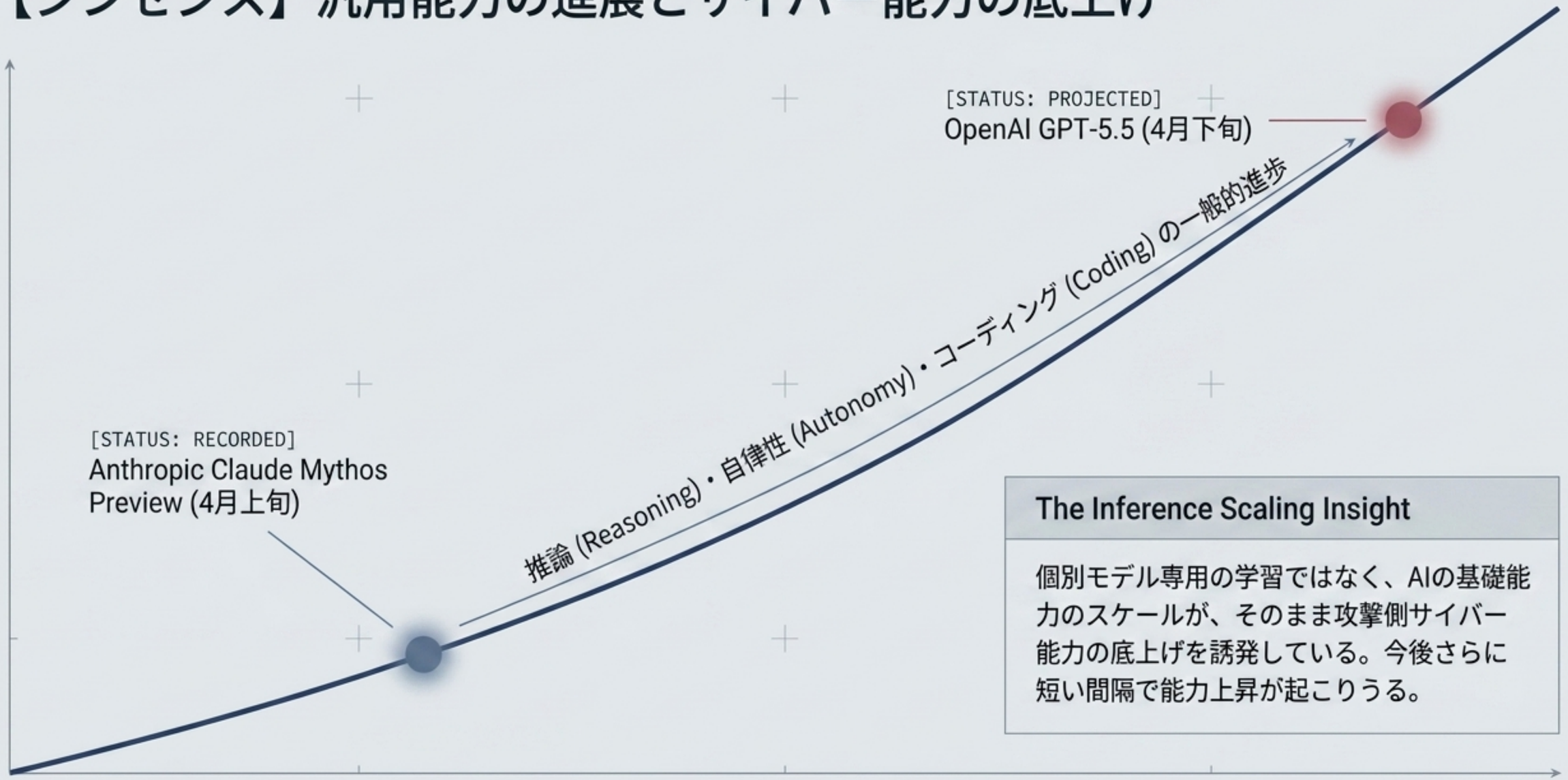
Real World System (現実環境)

[ROBUST]

[CONCLUSION]

評価結果を「強固に防御された本番システムへの成功確率」として一般化することは不可。監視・分離・検知を前提に読み解くべき。

【シンセシス】汎用能力の進展とサイバー能力の底上げ



The Inference Scaling Insight

個別モデル専用の学習ではなく、AIの基礎能力のスケールが、そのまま攻撃側サイバー能力の底上げを誘発している。今後さらに短い間隔で能力上昇が起こりうる。

政策と運用のパラダイムシフト

[STATUS: DEPRECATED]

Past: 壁を作る (Block & Ban)

AIの公開を全面的に禁止し、能力を隠蔽する。

[STATUS: IMPLEMENTING]

Future: 統制と防御の強化 (Controlled Access & Arming Defenders)

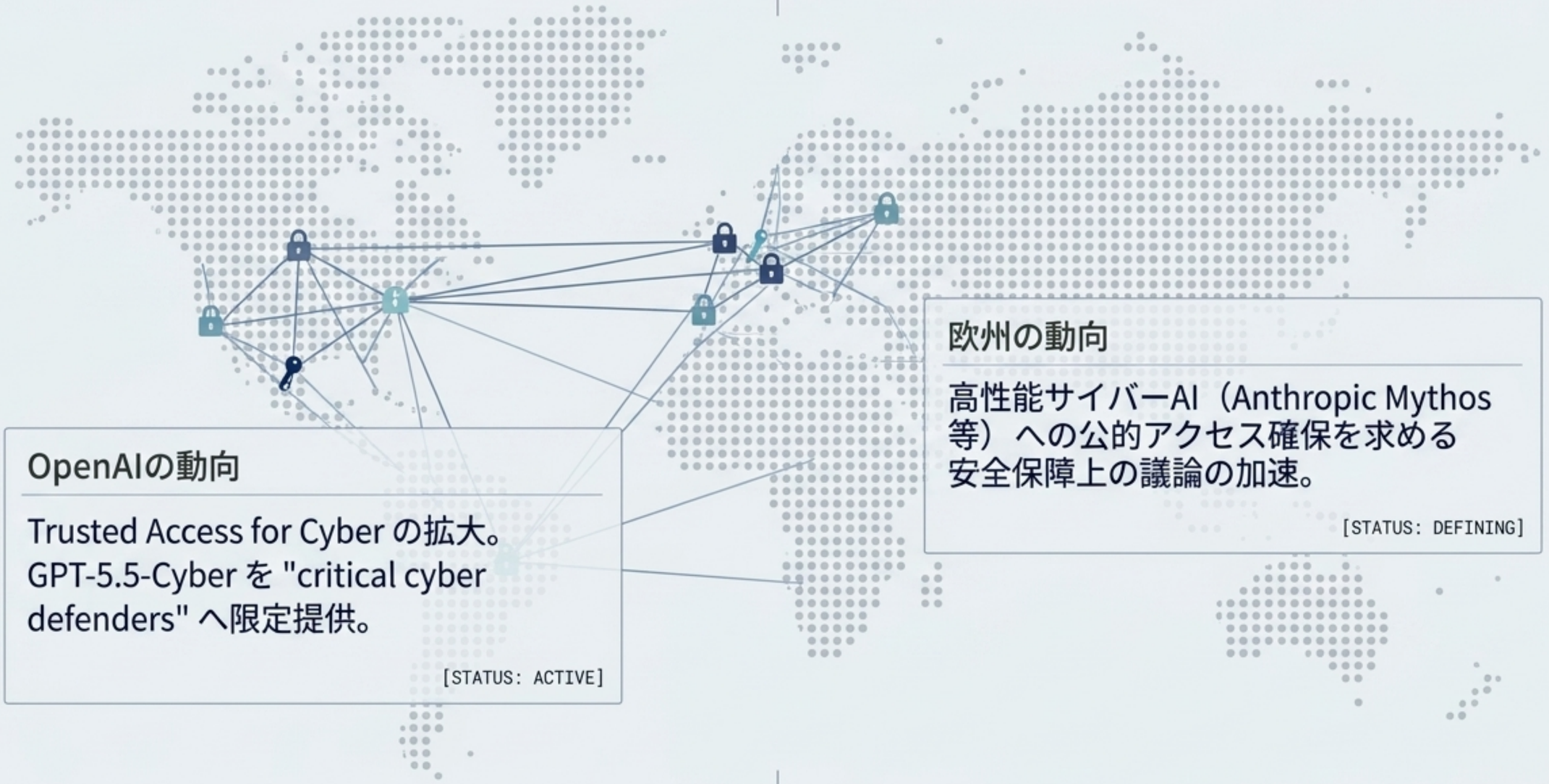
段階的アクセス制御（Trusted Access）。より強力なモデルを防御側組織に優先配分し、迫り来る "vulnerability patch wave"（脆弱性発見の波）に備える。

日本企業向け実務アクションマップ

	防御基盤の強化 (Defense)	AIの積極利用 (AI Usage)
経営層 (Management)	**1 パッチ適用の経営課題化** 月例パッチや属人的対応からの脱却。 [STATUS: DEFINED]	**2 社内LLMのサイバー用途監査** 一般設定と特権アクセス（permissive access）の明確な分離と監査。 [STATUS: DEFINED]
現場 (Field Operations)	**3 ログ設計のAIノイズ対応** 高速化する攻撃やAI生成ノイズに耐える検知基盤の再設計。 [STATUS: ACTIVE]	**4 防御ユースの試験導入** RE、脆弱性調査、マルウェア解析を自動化する防御テストを限定環境で開始。 [STATUS: ACTIVE]

国際的な地政学・アクセス統制の未来

The Future of Access Control



OpenAIの動向

Trusted Access for Cyber の拡大。
GPT-5.5-Cyber を "critical cyber defenders" へ限定提供。

[STATUS: ACTIVE]

欧州の動向

高性能サイバーAI（Anthropic Mythos 等）への公的アクセス確保を求める安全保障上の議論の加速。

[STATUS: DEFINING]

「モデル性能」の議論から、「誰が、どの条件でモデルを統治・利用できるか」という安全保障の論点へ完全シフト。

未解決の問いと今後の注視点

[UNVERIFIED]	AISIによる完全版報告書および技術補遺PDFの公開。
[UNVERIFIED]	企業ネットワーク完遂「2/10」修正値の、OpenAI System Cardへの公式反映タイミング。
[UNVERIFIED]	評価対象チェックポイントと、一般公開API・Trusted Access間の厳密な構成差分の開示。
[UNVERIFIED]	独立した第三者機関による、同一レンジ・同等条件での再現実験結果。

これらの未確定要素がクリアになるまで、評価の断定は避け、防御姿勢の強化を優先せよ。