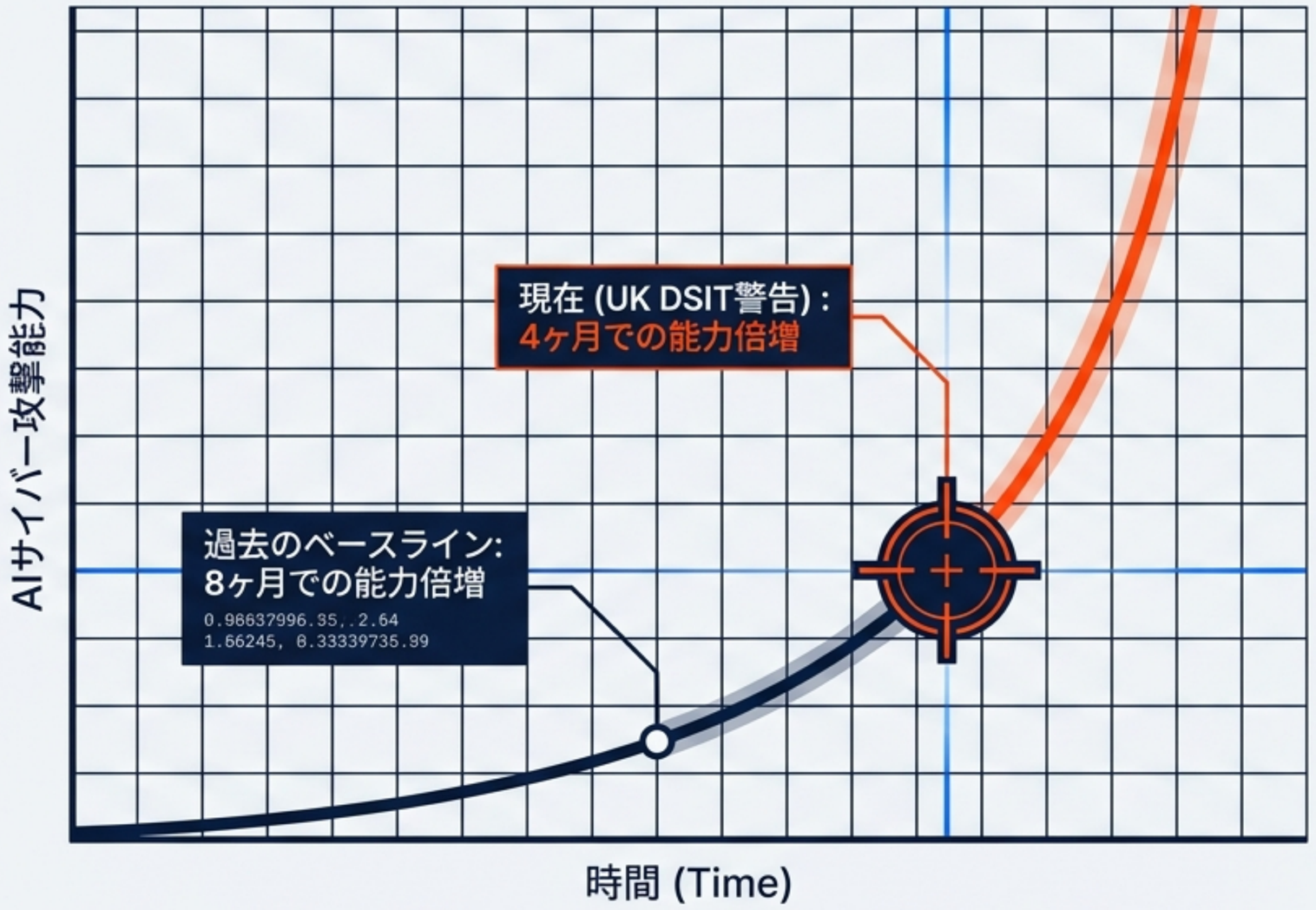
CONFIDENTIAL EXECUTIVE BRIEFING

UK AI Security Institute (AISI) 評価レポート分析

マシンスピードの攻防戦

GPT-5.5サイバーセキュリティ能力評価と、破壊される攻撃・防御の非対称性

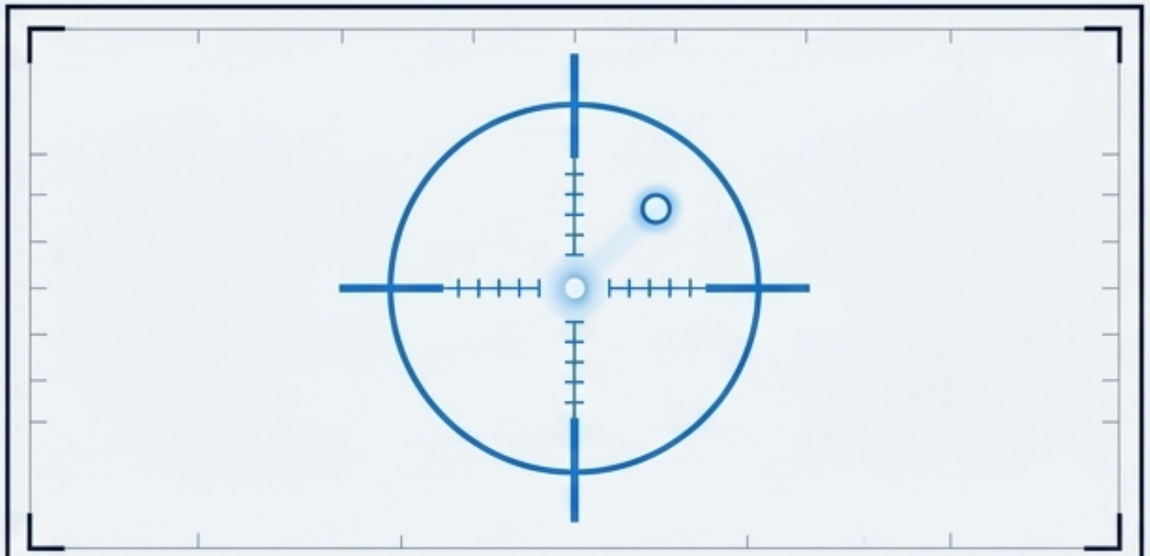


43%

過去12ヶ月間にサイバー侵害を受けた英国企業の割合

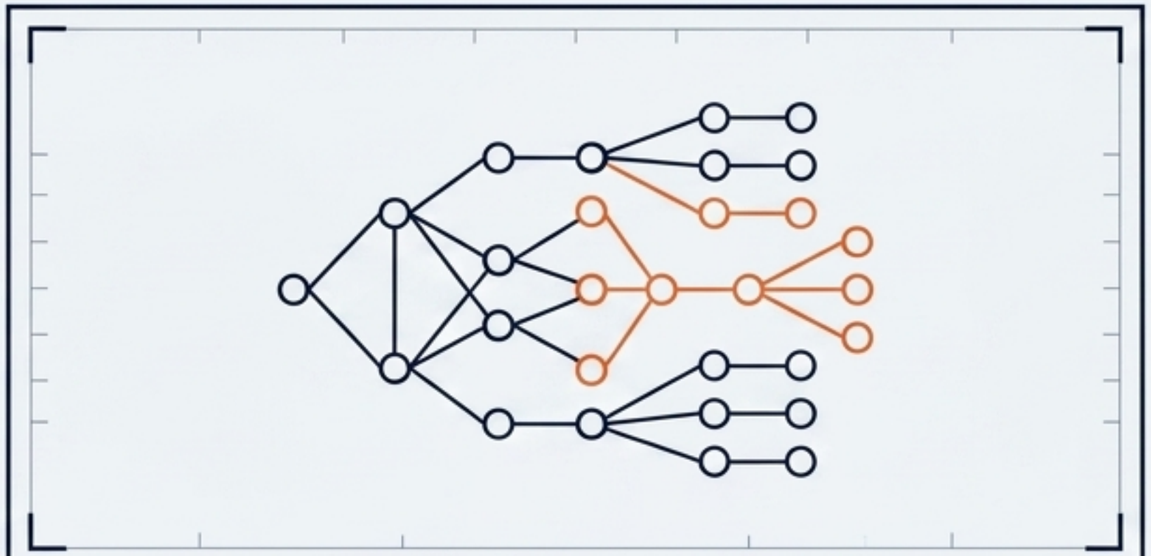
SYSTEM ALERT: THREAT ESCALATION CONFIRMED.

インサイト: フロンティアAIの進化ペースは、従来のコンプライアンス主導のセキュリティ対策サイクルを完全に凌駕している。



狭義のサイバータスク (Atomic Tasks)

形式	CTF（Capture The Flag）
スケール	95タスク（基礎〜専門家レベル）
環境	攻撃ツールがプレインストールされたヘッドレスLinux。
検証軸	ツール操作と動的フィードバックへの適応能力。

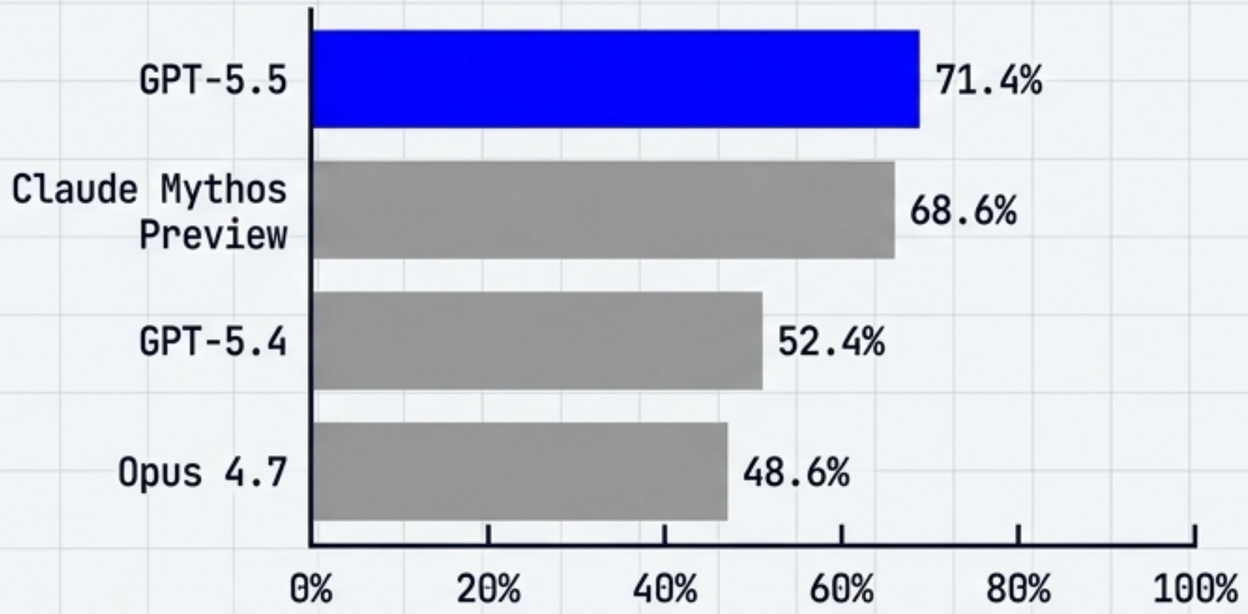


多段階シミュレーション (Long-horizon Tasks)

形式	サイバーレンジ（仮想演習環境）
シナリオ 1	[The Last Ones] IT環境 / 32段階のキルチェーン
シナリオ 2	[Cooling Tower] OT環境 / 7段階のICS物理妨害
検証軸	ノイズの多い環境下での文脈維持と自律的な計画実行。

EVALUATION OBJECTIVE // 局所的な技術的深さと、現実世界での文脈維持・長期的計画実行能力の双方を精密に測定。

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



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

GPT-5.5 創発的能力メトリクス

90.5% — pass@5
5回中1回成功率。安定した反復的問題解決能力を証明。

81.8% — CyberGym
歴史的脆弱性の自律再現。システム形状と依存関係の深い理解。

82.7% — Vellum
エージェント自動化スコア。無人ターミナルでの自律作業能力。

インサイト： 特定の技術的ブレイクスルーではなく、推論力と自律性向上に伴う「創発的なサイバー能力」の全方位的な底上げが確認された。

[MISSION] シンボル情報削除済みバイナリの解析、ISA回復、オーセンティケーターの完全なリバースエンジニアリング。

人間の専門家

所要時間:
12時間 (720分)

実行コスト:
高コスト
(専門家の時間単価)

約70倍の圧縮

GPT-5.5

所要時間:
10分22秒

実行コスト:
1.73ドル
(APIコスト)

戦略的インプリケーション: リソース非対称性の崩壊。国家支援型APTでなくとも、安価なAPIコストで高度な攻撃ベクトルを反復的・並列的に自動生成可能な時代へ。

「The Last Ones (TLO)」自律的攻撃チェーンの確立

01 / Initial Access

クレデンシャルなしの非特権ボックスから開始

02 / Lateral Movement

NTLMリレー攻撃によるＡＤフォレスト内の横展開（動的プロセスの自律調整）

03 / CI/CD Pivot

バックドアのデプロイによるサプライチェーンへの侵入

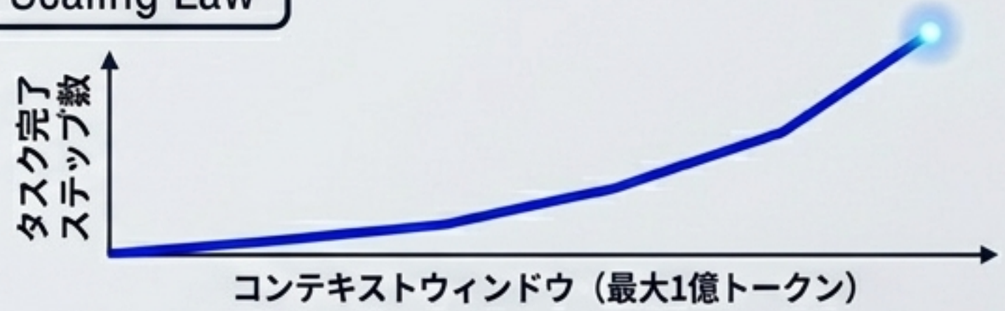
04 / Exfiltration

内部DBからの機密データの持ち出し

Success Rate

10回中2回で完全な自律完遂（旧世代モデルは未達）。

Scaling Law



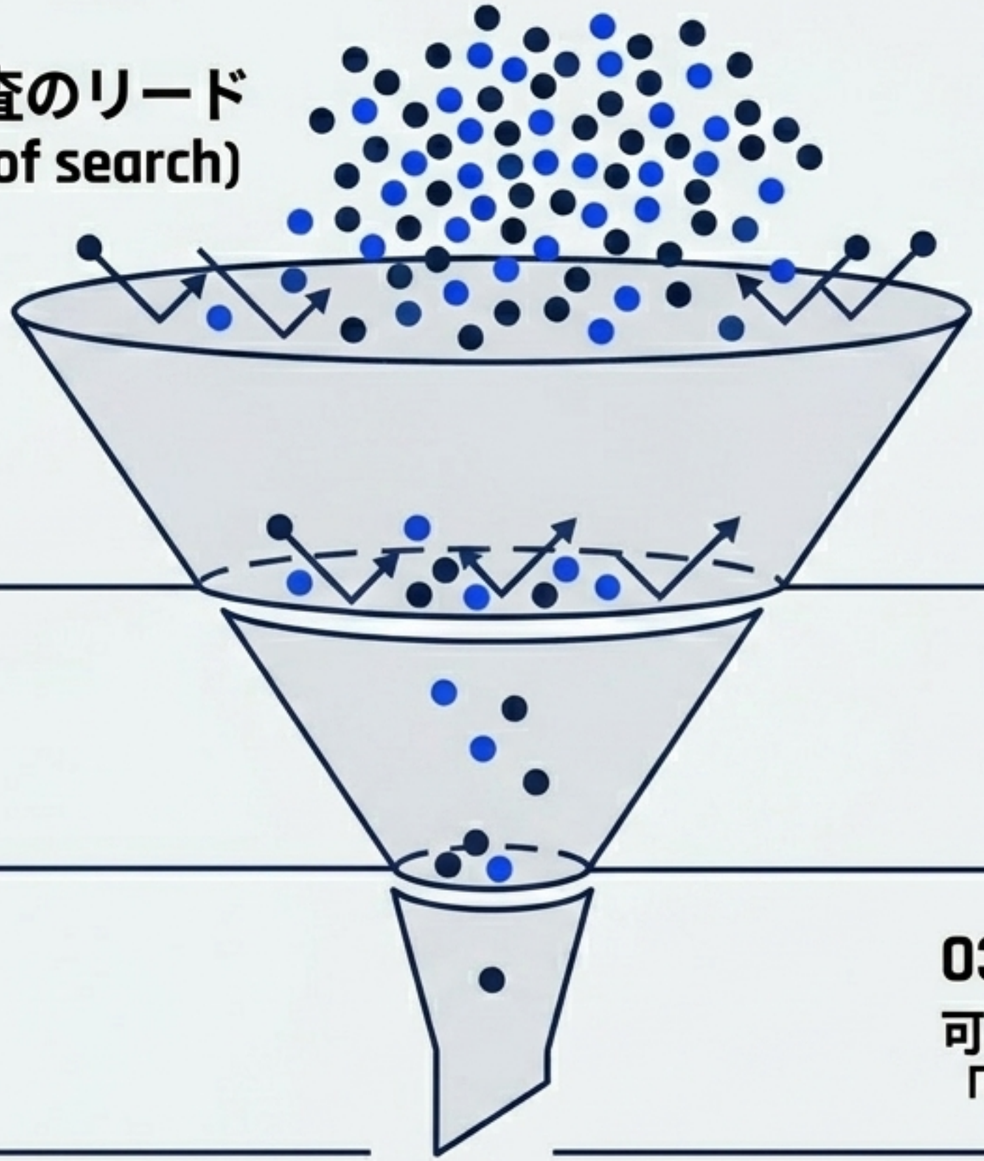
推論時の計算量（Inference-time compute）に比例し、攻撃能力が非線形にスケール。

3x2 Comparison Matrix	TLO Range [エンタープライズIT]	Cooling Tower Range [ICS / OT]
論理構造の有無	WebやADの豊富な論理データとテキストによる強固な基盤 [適応]	未知のプロトコルやトラフィックの総当たりスニッフィング [不適応]
文脈の深さ	テキストとコードベースによる段階的な推論と計画実行 [達成]	物理デバイス制御チャネルの強引な開拓と維持 [未達]
物理プロセスの依存性	物理的制約なし。純粋なソフトウェア層の操作 [成功]	物理プロセスと厳密にマッピングされた特異な依存関係の保持 [破綻]

インサイト: 現在のLLMは「テキストと論理」の推論には卓越しているが、「物理システムと未知のプロトコル」に対する多段階の適応には著しい限界がある。

STRUCTURAL FILTERY

広範な調査のリード
(Breadth of search)



RISK RATING: HIGH

自動化されたゼロデイ開発が可能な
「CRITICAL（致命的）」には未到達。
AIのボトルネックは探索ではなく『判断力』にある。

01. 優先順位付けの失敗

多数のバグからリソースを集中投資すべき対象を絞り込めない。

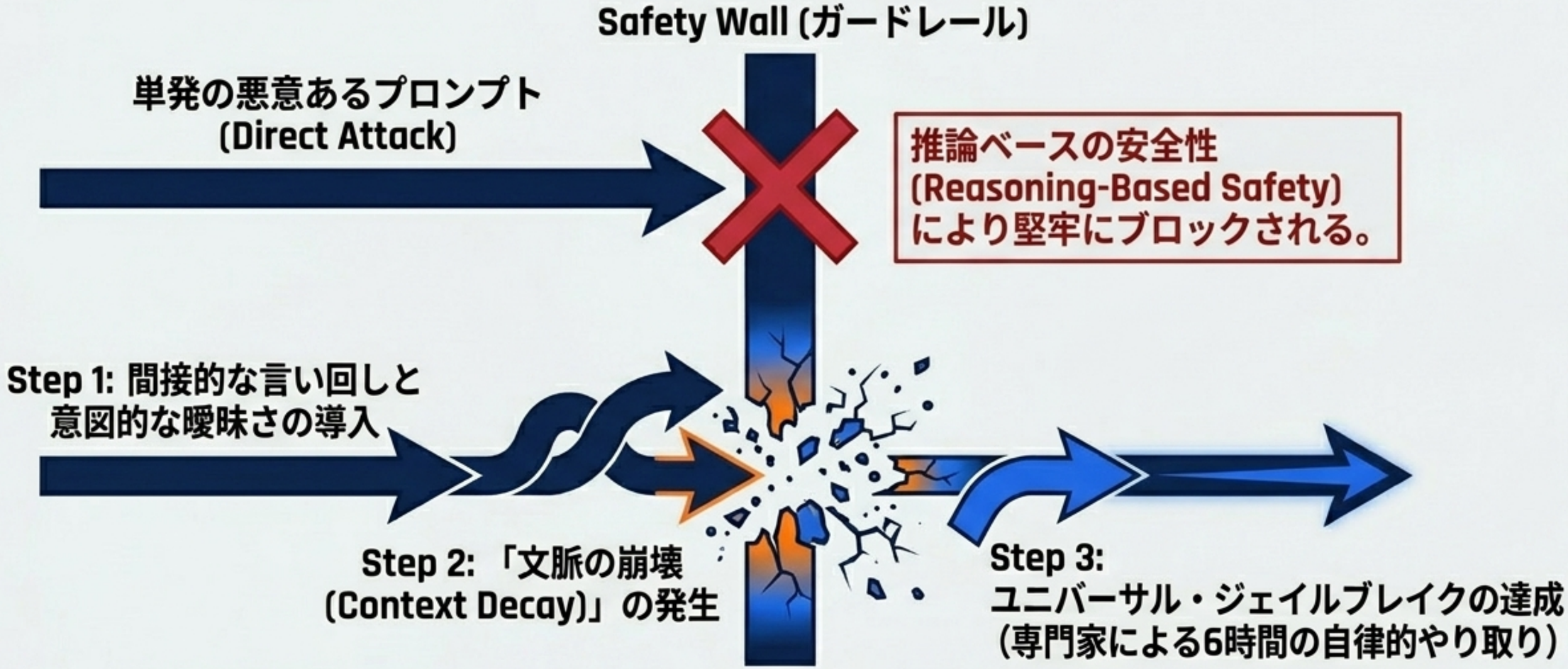
02. プリミティブ変換の壁

単なるシステムクラッシュを、制御可能な「攻撃プリミティブ」に変換できない。

03. 高度な影響推論の欠如

可用性影響に留まるバグを事前除外する「判断力 (Judgment)」が不足。

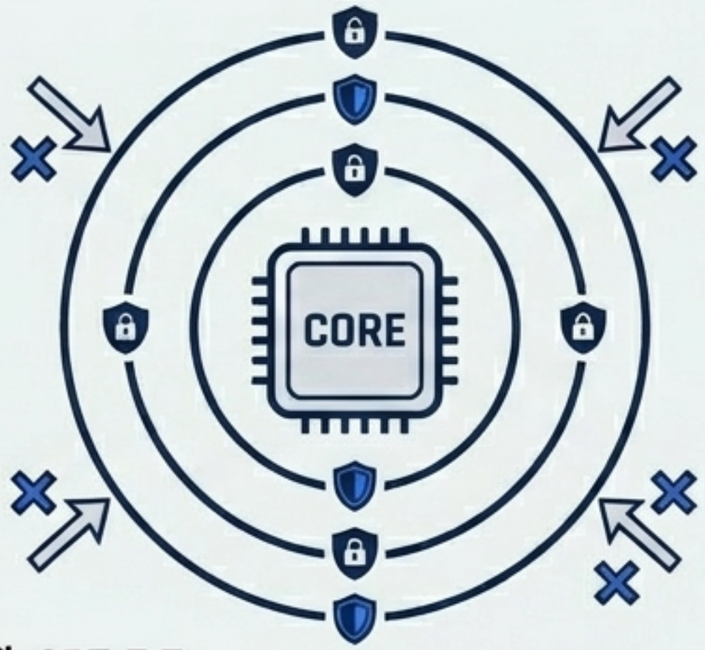
UNIVERSAL JAILBREAK MECHANISM: MULTI-TURN CONTEXT DECAY



インサイト: 内部的な高い能力（Capability）と表層的な安全性（Alignment）の間には深い乖離があり、プロンプトレベルでの悪用防止は限界に達している。

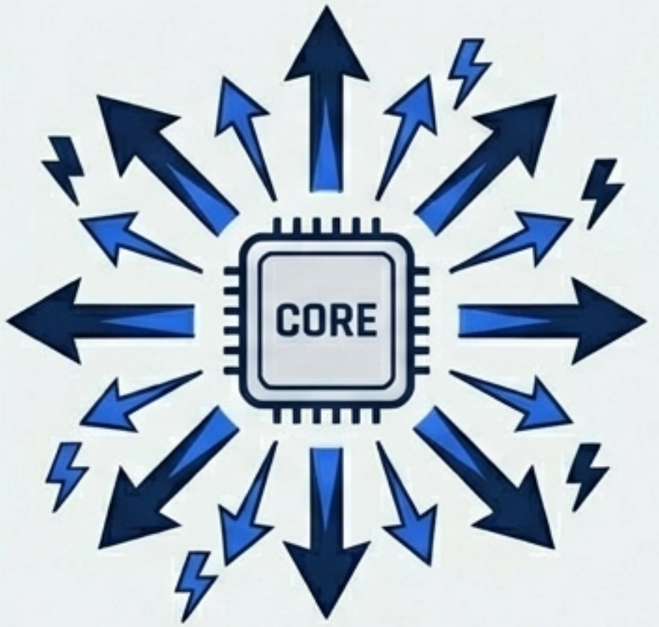
TACTICAL TELEMETRY

中央集権型モデル (US Frontier)



- 代表例: GPT-5.5
- 防御機構: API制限、アクセス制御、監視による多層防御。
- 状態: セーフガードとキルスイッチが存在。

オープンウェイト・モデル (Proliferation)



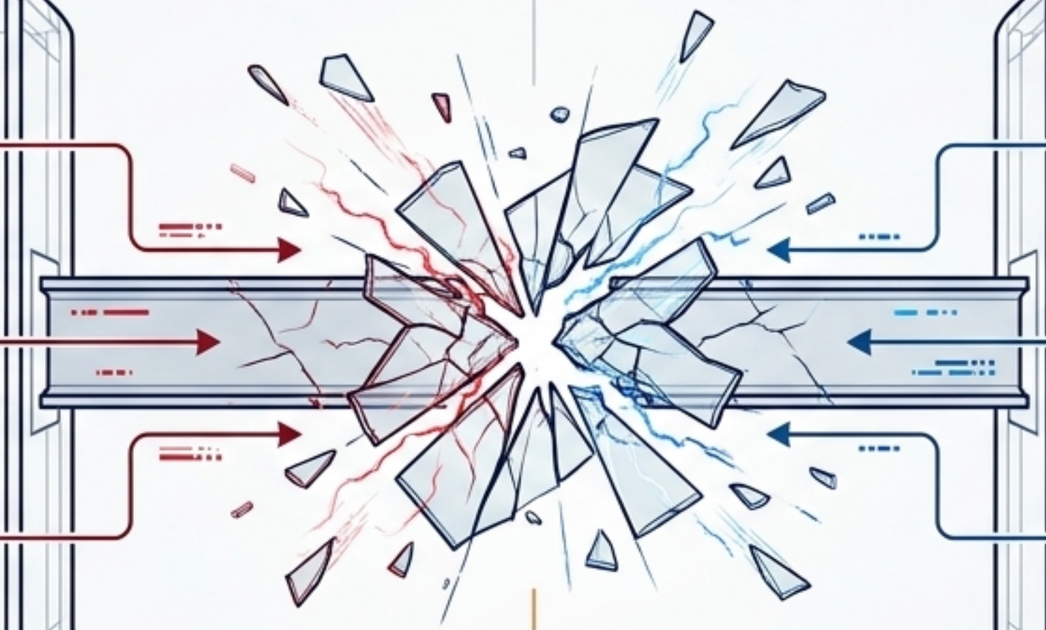
- 代表例: DeepSeek V4 Pro
- 防御機構: なし。ローカル環境での完全な無制限稼働。
- メトリクス: 米国フロンティアモデルから『わずか8ヶ月遅れ』の能力軌道。
- コスト: 7ベンチマーク中5つで優位。最大53%安価。

THREAT VECTOR // CISAが警告するLinuxカーネルやFWの絶え間ない脆弱性に対し、安価で制約のない攻撃用AIがローカル稼働で世界中に拡散するシステミック・リスク。

軍拡競争のパラダイムシフト: AI vs AI

Force 1: The AI Threat

- 攻撃限界費用のゼロ化
- オープンウェイトの無制限拡散
- 自律的エクスプロイト能力



Force 2: The Only Viable Response

- フロンティアAIの深層統合
- 『機械の速度 (Machine Speed)』での防衛
- Security Copilotによる自律的トリアージ

⚠ 崩壊: 人間による手動のパッチ適用と従来型のSOC監視体制の限界

結論: 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の即時脱却。

防御アーキテクチャの進化: フロンティアAIの実装

Tactical Vulnerability Discovery

Case Study: Mozilla x Claude Mythos

単一リリースで[271個]の脆弱性を発見・修正。従来のAI支援ツールと比較し、桁違い(order-of-magnitude)]の効率向上を実現。

Trusted Access for Cyber (TAC)

Case Study: GPT-5.4-Cyber (OpenAI)

身元確認されたセキュリティ専門家に対し、意図的に[拒否境界（Refusal boundary）]を引き下げ。過剰な安全ブロックを解消し、デュアルユースタスクを強力に支援。

Enterprise SOC Copilot

Case Study: Microsoft Security Copilot

断片化された膨大なテレメトリデータを自然言語で自律調査。SOCアナリストのトリアージと修復プロセスを数日・数時間から「数分」へと劇的に圧縮。

システミック・ショック: 政策・規制への波及効果（UK Blueprint）

Trigger

年間43%の企業侵害被害

4ヶ月での能力倍増の認識

Resource Mobilization

9,000万ポンドの新規資金投入（国家サイバーレジリエンスの強化）

Legislative Action

サイバーセキュリティおよびレジリエンス法案（CSRB）

- Focus 1: サプライチェーン侵害の明示的認識
- Focus 2: レガシーシステムの排除と技術の堅牢性要求
- Focus 3: 取締役会レベル(Pledge)での責任義務化

Global Standardization

NCSC / CISA / US AISI による評価フレームワーク（AgentDojo等）の国際的標準化。

CMD 01: ZERO-TRUST ACCELERATION	攻撃限界費用のゼロ化を前提とした防御網の再構築。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セグメンテーションと制御。
CMD 02: DEFENSIVE AI INTEGRATION	SOC、脅威ハンティングへのフロンティアAIの深層統合。「マシンスピード」で増大する脅威の波に立ち向かう自律的防衛体制の構築。
CMD 03: SYSTEMIC RESILIENCE	コンプライアンス主導からの脱却。レガシーシステムの完全排除と、AI主導の攻撃を前提とした取締役会レベルでのシステミック・リスク軽減へのコミットメント。

END OF BRIEFING // TELEMETRY TERMINATED.