INTELLIGENCE BRIEFING | 2026.10.06

日本版AISI サイバー能力評価の戦略的解読

企業のAIガバナンスと知財・R&D実務への実践的示唆

調査基準日: 2026年10月6日 | 情報源: GPT-6 Astra レポート

本格始動の意味: ガイドライン策定から「自律的・実測評価機関」への進化（予算88億円体制へ）。

評価結果のノイズ除去: Claude 4.8等の高スコア（T1）は限定環境下での「兆候」であり、即時の脅威を意味しない。

明日からの実務: 「観測と制御」の権限設計が必須。知財・R&Dの現場における自律AIエージェントのリスク管理へ直結。

Gemini Notebook

Executive Insight: 秋の「本格始動」とは組織の発足ではなく、自ら実測・公表する「評価機関」としての活動具体化である。

過去の課題
- 知見の集約主体
- 研究開発資源の不足（2026年4月内閣府指摘）

機能拡張
専用試験環境での自律的評価

予算



※恒常的な運営費とは区別

人員



Gemini Notebook

AIの出力・判断（Output）

誤情報、有害情報、情報漏洩

従来の主対象

AIへの攻撃耐性（Defense）

不適切な操作、レッドチーミング

既知の攻撃の整理

AI自身が実行できる行為（Autonomy）

ツール使用、高度な作業、サイバー能力

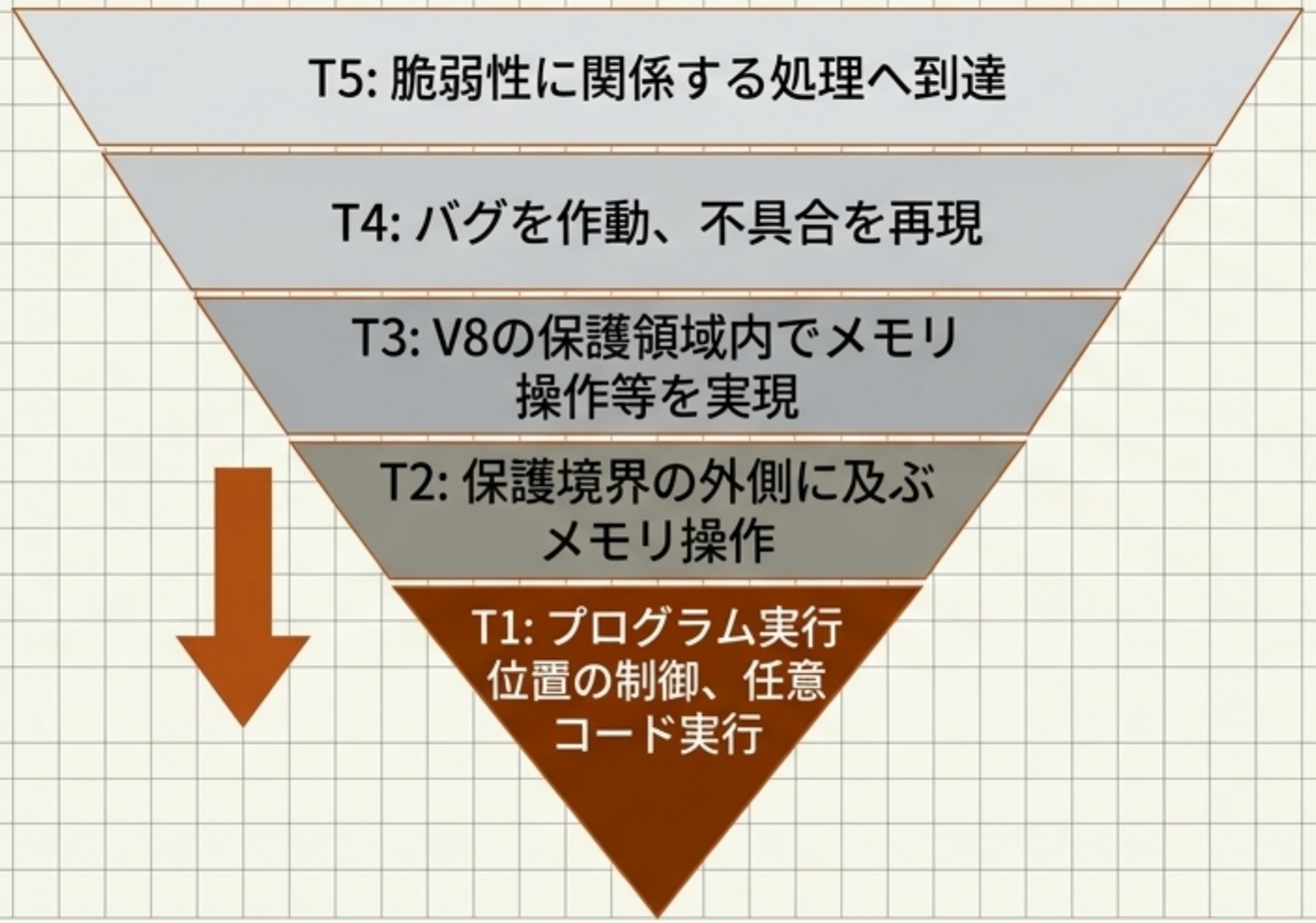
今回の評価範囲（エージェント評価）

Executive Insight: 今回の評価は「AIを攻撃から守る」のではなく、「AIが自律的に攻撃を実行する能力」を測るものである。

Gemini Notebook

Executive Insight: ベンチマークの到達深度は、直ちに現実の企業ネットワークへの侵入成功を意味するものではない。

Penetration Depth Funnel



Reality Check Box

現実の脅威とのギャップ（Noise Filter）

ExploitBenchは「既知のV8脆弱性」を対象とする。
未知の脆弱性発見能力や、企業ネットワーク全体への侵入能力を直接測るものではない。

Gemini Notebook

条件（Context）

通常提供時のセーフガードを「解除」。
最大300ターン・5時間の試行許可。

到達点（Signal）

T1レベルの「**pc_control**（実行位置の制御）」を1件獲得。
前モデルから明らかに上位段階への到達が増加。

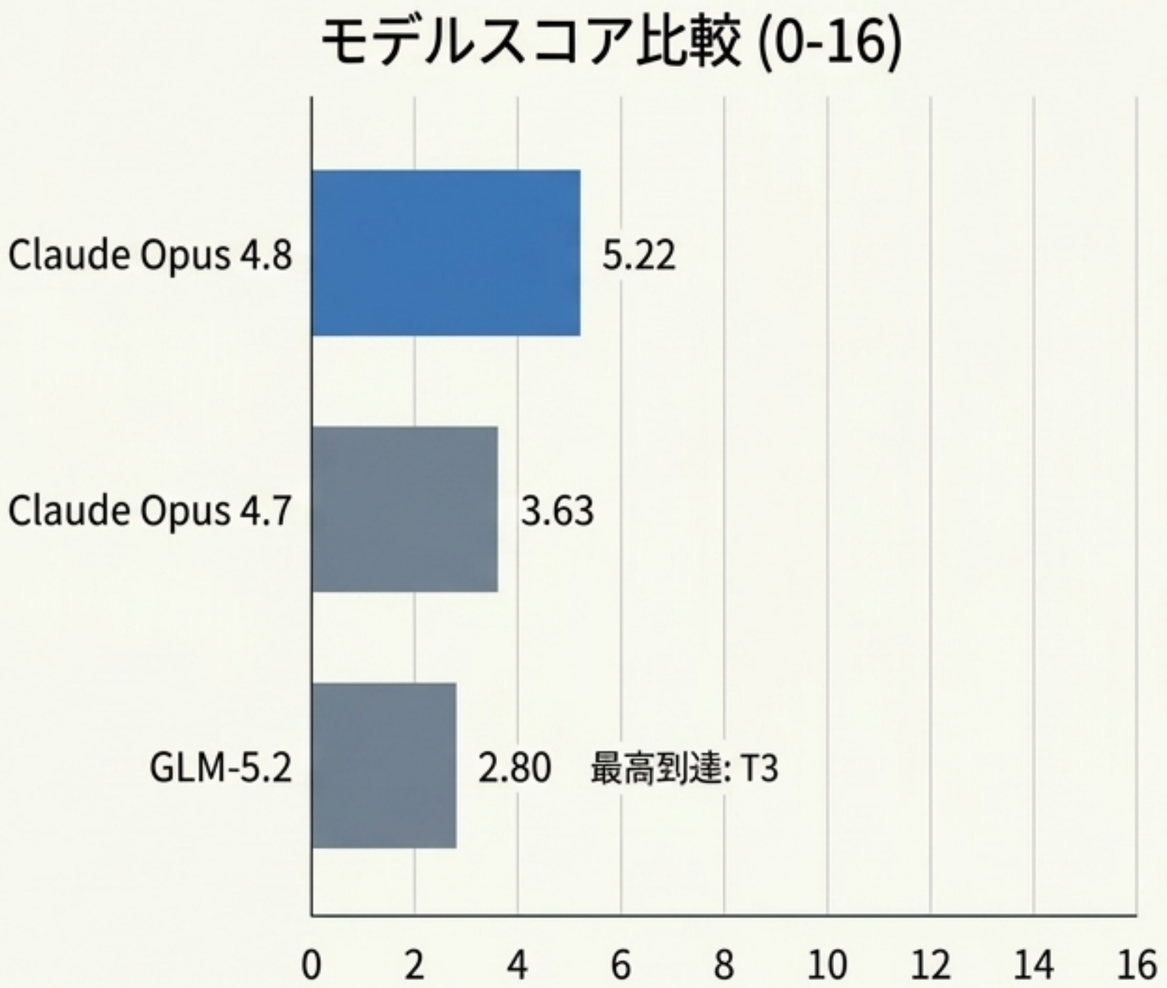
限界（Limits）

最高脅威である「ace (任意操作)」には未達。
追試で pc_control の再現性なし（想定された段階を順番に進まずに高能力が観測された）。

Executive Insight: 通常利用でのリスク不存在の証明でも、確実な脅威の断定でもない。「限定条件での能力の兆候」として扱うべきである。

Gemini Notebook

Executive Insight: 技術的能力（スコア）と安全制御（ガードレール）は別次元の評価軸である。



Safety Control Alert

安全制御に関する特記事項

GLM-5.2では、この評価中「安全上の拒否応答」が確認されなかった。

Translation to Biz

正当な研究環境で拒否しなかったという観察であり、実世界での悪用に対する総合的な拒否性能を断定するものではない。単一モデルの結果をオープンモデル全体へ一般化しないこと。

Gemini Notebook

1.
Model 3はCriticalを「0」と判定。しかし他モデルは3〜4件報告。

モデル	Critical	High	Medium	Low
Model 1	4	5	6	0
Model 2	3	5	4	1
Model 3	0	5	6	2

2.
候補の細分化や重大性の解釈がモデルごとに異なる（例：Model 1の細分化傾向）。

Heatmap Table

複数モデルの共通指摘は「優先度高」だが、単独指摘を捨てる理由にはならない。
知財調査でも「件数の多さ＝見落としの少なさ」ではない。

Executive Insight: AIの指摘件数だけで優劣は決まらない。1つのモデルしか指摘しなかった重大リスク（死角）を見落としてはならない。

Gemini Notebook

Executive Insight: 日米英の評価の違いは「矛盾」ではなく、「対象・条件・時点」の違いによるスナップショットである。

	日本 (AISI)	英国 (UK AISI)	米国 (NIST/CAISI)
対象 (Target)	V8既知脆弱性	模擬ネットワーク上の複数段階作業	後継モデル(GLM-5.3)の比較評価
主な問い (Core Question)	どの技術的段階まで到達するか	長い作業の遂行・モデル間の差	最先端モデルとの継続的な差
読み取れること (Key Insight)	特定の高度な攻撃学技術の限界	より広いサイバー能力の進展	最先端モデルとの比較
時点 (Timing)	2026年6-7月時点	-	2026年9月

Gemini Notebook

Strategic Intelligence Briefing

評価する点（+）

- モデル名と結果の公開
- 実行記録の解釈
- ガイド/ツールの組合せ

残る課題（-）

- 複数回試行（ばらつき）の報告
- 実利用実績の指標化
- 評価の独立性

今後確認すべき5つの課題

1. **設定の特定**
（バージョン、推論設定、安全制御の有無）
2. **結果の安定性**
（試行回数、信頼区間、まぐれ当たりの頻度）
3. **実際の利用との対応**
（通常時 vs セーフガード解除時）
4. **試験対象の広さ**
（外部ツール、長時間作業、複数エージェント）
5. **公表の速さ**
（新モデルへの迅速な追試）

Executive Insight: 評価機関としての実測への進展は高く評価できるが、企業が意思決定に利用するには「統計的信頼性」と「公表スピード」が課題となる。

Gemini Notebook

Executive Insight: 単なる性能試験機関ではなく、産業ごとのAI実装を支える「包括的なツールセット」が整備されつつある。

ガバナンス層（Governance）

Chief AI Officerガイド：責任者、組織、調達、運用、監査 (2026.3)

評価・実践層（Evaluation）

評価観点ガイド v1.20：「観測と制御」を追加。導入・更新時の基礎

既知の攻撃と影響：試験シナリオ作成支援

ツール・環境層（Technical Tools）

aisev（AIセーフティ評価環境）：OSS公開。自動レッドチーミング（2025.9）

※政府の公式認定を意味するものではない

業界特化層（Industry Specifics）

事業実証WG：ヘルスケア、ロボティクス、教育、ジオAI等の分野別展開

Gemini Notebook

Diagnostic Table: 実務リスクと評価事項 (Use Case & Evaluation Metrics)

実務リスク (Use Case)	評価事項 (Evaluation Metrics)
【特許SDI/監視】	検索の再現性、重要案件の見落とし、根拠公報への追跡可能性（ハルシネーション対策）。
【明細書作成】	未公開技術情報の送信先、参照資料の取り違え、根拠のない記述。
【エージェント利用】	閲覧権限、外部文書からのプロンプト インジェクション耐性、情報持ち出しリスク。
【出願・案件自動化】	書き込み権限、重要操作の人間による承認(Human-in-the-loop)、監査記録。

Executive Insight: どのモデルを使うか以上に、「何を読ませ、どのツールを使わせ、どこまで自律実行させるか」の権限設計が知財競争力を左右する。

Gemini Notebook

2026-2027 Roadmap: Timeline & Scenario Board

2026-2027 Future Roadmap

① 個別評価の継続

個別モデル評価公表の継続・迅速化。更新スピードと対象範囲（オープン/クローズド）の拡大が焦点。

② 自律エージェントの総合評価

単体モデルの性能から、「ツール＋権限＋記憶」を組み合わせた自律エージェント全体の検証へ移行。

③ 業界別テスト環境

事業実証WGを通じた、業界特化型の評価データ・試験環境の具体化。

④ 調達要件への発展

ガイドや評価結果が、将来的な「企業のAI調達・監査の要求仕様」や「相互承認の枠組み」へ発展する可能性（注視が必要）。

Gemini Notebook

THE CAIO ACTION PLAN (Strategic Directives)

評価の「文脈」を問う

単なるベンチマークスコアを鵜呑みにせず、試験条件（セーフガードの有無、試行回数、再現性）を必ず確認し、リスクを過大・過小評価しない。

「自律性」のリスクを管理する

優秀なモデルであっても、ツール利用やファイル操作の権限（観測と制御）は厳格に設計する。エージェント化によるリスク増大に備える。

AISIエコシステムを「実装」する

提供されるガイド（v1.20）やOSS評価環境（aisev）を自社の調達・監査プロセスに組み 込み、AIリスク評価能力を早期に内製化する。

Gemini Notebook