日本版 AISI の組織・活動と AI サイバー能力評価

Claude Opus 4.8・GLM-5.2 の検証結果、評価の限界、企業・知財実務への示唆

調査基準日：2026 年 10 月 6 日

GPT-6 Astra

文中番号引用・巻末参考文献付き

調査概要

日本版 AISI は、AI 安全性のガイドや評価基盤を整備する活動に加え、個別の AI モデルを実際に動かし、その能力と限界を検証する機能を強化している。今回のサイバー能力評価は、その進展を示す具体的な成果である。AISI は 2024 年 2 月に設立されており、2026 年秋の「本格始動」は組織の発足ではなく、自ら実測して結果を公表する評価機関としての活動の具体化と理解するのが適切である。[1][4][7]

本報告書は 2026 年 10 月 6 日時点で確認した公開資料に基づく。ご提示の検証ノート 3 件、内閣府の組織・予算資料、ExploitBench の原論文、英国・米国の公式評価を照合し、組織の役割、技術的な結果、評価の限界、企業・知財実務への示唆を整理した。事実の記載には文献番号を付し、独自の評価・提案・予測はその旨を明示する。

調査上の制約として、YouTube 動画の音声・字幕全文と、提示された X 投稿の全文は取得できていない。動画に対応する解説記事と投稿の検索結果は確認したが、技術的な判断は公式ノート、原論文、政府資料に基づく。動画・SNS の発言全体を逐語的に検証したものではない。[25]

1. 日本版 AISI の役割と組織

正式名称は「AI セーフティ・インスティテュート」、英語では AI Safety Institute である。日本における AI 安全性の中心的機関として、評価手法・基準の検討、知見の集約、国内外の連携等を担う。[1][2]

項目	確認できた内容
発足	2024 年 2 月 14 日。2023 年 12 月の設立表明を受けて発足。
設置場所	独立行政法人情報処理推進機構（IPA）内に事務局を設置。
所長	村上明子氏。公式サイトでは非常勤と記載。
連携体制	12 府省庁・5 関係機関が参画する横断的な体制。
主な協力機関	IPA、産業技術総合研究所、情報通信研究機構、国立情報学研究所、理化学研究所。
運営	関係府省庁等連絡会議と AISI 運営委員会を通じ、政府方針との連携や事業運営を検討。
事務局の構成	組織説明資料には戦略企画、評価リサーチ、安全検証の各グループが示されている。

出典：[1][2]。組織・役職は各資料の記載に基づく。AISI の守備範囲はサイバー攻撃対策に限らず、誤情報、有害な出力、公平性、プライバシー、説明可能性、データ品質、AI エージェントの制御等に及ぶ。[5]

制度上は、評価結果の公表と、政府による認証・承認を区別する必要がある。今回の検証ノートは、対象モデルの安全性や法令適合性を政府が認定・保証するものではないと明記している。適合性評価については別途、制度設計に向けた検討が行われている。[7][9][22]

2. 機能強化の背景、人員、予算

2026 年 4 月 7 日の内閣府資料は、当時の AISI について、技術評価の研究開発資源が不足し、国立研究機関の知見を束ねる機能や、自らモデルを評価する能力が十分ではないと評価していた。機能強化は、この課題認識を背景としている。[3]

項目	公開資料の記載	解釈上の注意
職員数	2026 年 4 月資料で 31 人。併任を含む。	10 月現在の確定人数ではない。

項目	公開資料の記載	解釈上の注意
当面の増員	約 30 人から 60 人体制への移行を推進。	整備方針であり、達成済みとは限らない。
より大きな目標	2025 年 12 月の総理指示で 200 人体制を目指す。	目標と実員を区別する。
機能強化の財源	令和 7 年度補正予算で合計 88 億円を措置。	恒常的な年間運営費とは区別する。

出典：[3]。予算の用途には、評価手法、専用の試験環境・計算資源、AI エージェント向けの評価ツール、ロボティクス関連の研究等が含まれる。過去の「令和 6 年度補正 3.8 億円」と、この機能強化の補正予算を、対象範囲を揃えずに比較することも避けるべきである。[3]

【評価】今回の公表は、機能強化の方針に対して実際のモデル評価という成果が現れた点で意義がある。一方、公開資料だけでは、現在の常勤研究者数、設備の稼働状況、作業の内製・外注の分担までは確認できない。評価機関としての活動が進展したと評価できるが、組織整備の完了を示すものとは判断できない。

3. AISI が評価する三つの対象

AI 安全性の議論では、AI の出力・判断、AI システムへの攻撃耐性、AI 自身が実行できる行為を分けると、今回の評価の位置付けが明確になる。以下は AISI の活動を理解するための整理である。[5][6][7]

評価の対象	主な問い	今回との関係
AI の出力・判断	誤情報、有害情報、偏り、情報漏洩をどの程度抑えられるか。	従来からの評価ガイドの主要な対象。
AI システムへの攻撃耐性	外部の入力やデータによって、AI が不適切に操作されないか。	レッドチーミング、既知の攻撃の整理。

評価の対象	主な問い	今回との関係
AI 自身が実行できる行為	ツールを使い、どの程度高度な作業や危険な行為を実行できるか。	サイバー能力評価、エージェント評価。

AI を攻撃から守る評価と、AI が攻撃に使われた場合の能力を測る評価では、必要な試験が異なる。AISI の活動は、その双方と社会実装に必要な評価を含む。

4. ExploitBench は何を測るのか

ExploitBench はカーネギーメロン大学の研究者らが開発したベンチマークで、Chrome 等に使われる JavaScript エンジン V8 の既知脆弱性を対象とする。能力を 16 項目に分解し、実際の動作によって判定する。AI が「成功した」と回答しただけでは成功としない点が特徴である。[8]

段階	おおまかな意味
T5	脆弱性に関係する処理へ到達する。
T4	バグを作動させ、不具合を再現する。
T3	V8 の保護領域内で、メモリ操作等を実現する。
T2	保護境界の外側に及ぶメモリ操作等を実現する。
T1	プログラムの実行位置の制御、任意コード実行等を実現する。

出典：[8]。数字が小さいほど、システムへの制御が深まる。原論文は修正情報がある既知脆弱性を扱う設定であり、未知の脆弱性を発見する能力をそのまま測る試験ではない。

【解釈】スコアを現実の企業への侵入成功率とみなすことはできない。また、V8 の保護境界を越える能力と、組織のネットワーク全体に侵入して活動する能力も異なる。ベンチマークの到達段階、取得した能力項目、現実の被害に至る一連の行動を区別する必要がある。

5. Claude Opus 4.8 の評価結果と意味

AISI が確認したのは、前モデルからの能力向上と、高い到達結果の解釈の難しさである。主な条件と結果を以下に整理する。[7]

確認事項	公式ノートの内容
実施・公表時期	2026 年 6 月に評価し、10 月 2 日に公表。
対象・試行	既知脆弱性 41 件。各 1 試行、最大 300 ターン・5 時間。
重要な条件	通常提供時のリアルタイム・サイバーセーフガードを解除して実施。
前モデルとの比較	比較可能な 40 件中、上位段階 15 件、同段階 23 件、下位段階 2 件。
特徴的な結果	1 件で実行位置を制御する能力「pc_control」を獲得。
その限界	任意操作に当たる「ace」には到達せず、追試でも pc_control は再現しなかった。

当該事例では想定された段階を順番に進まずに高い能力項目が観測され、その成果を後続の操作へ発展させられなかった。AISI は、スコアだけで判断せず、実行記録と再現性を調べる必要性を指摘している。[7]

【評価】報告の価値は、「何ができたのか」「何ができなかったのか」「もう一度できるのか」を分解している点にある。通常利用と異なる条件のため、一般の Claude 利用でも同じ行動が可能だとは結論付けられない。他方、追試で再現しなかったことだけで、その能力やリスクの不存在を証明したことにもならない。限定条件での能力の兆候として扱うのが適切である。

6. GLM-5.2 とオープンモデルへの能力の拡がり

2026 年 7 月に実施された比較では、V8 の 41 件のタスクを各モデルで 1 回ずつ実行した。平均スコアは次のとおりである。[9]

モデル	平均スコア（0〜16）

モデル	平均スコア（0〜16）
Claude Opus 4.8	5.22
Claude Opus 4.7	3.63
GLM-5.2	2.80

GLM-5.2 の最高到達段階は T3 であった。別途、10 件について行った助言・指示（ナッジ）の有無の比較では、平均スコアはどちらも 3.0 で、高度な段階への改善は限定的だった。また、この評価中、GLM-5.2 では安全上の拒否応答が確認されなかったと報告されている。[9]

【解釈】技術的な能力と、安全制御の挙動は異なる評価軸である。この試験での GLM-5.2 の能力は比較対象を下回ったが、正当な研究環境で拒否しなかったという観察だけでは、実際の悪用依頼に対する拒否性能の総合評価にはならない。単一モデルの結果をオープンウェイトモデル全体へ一般化することもできない。

AISI は、モデル更新や実行環境の改良によって能力が変化し得るため、継続評価が必要としている。同ノートには T1 に到達したタスクがあるとの記述もあるが、T1 に属する一部能力の取得と任意コード実行の達成は区別すべきである。また、6 月実施の Claude ノートと 7 月実施の比較ノートを、同一試験の結果として混ぜないことが重要である。[7][9]

7. 7 月 8 日の匿名モデル比較が示す実務上の注意

同じソフトウェアに対して通常利用できる三つのモデルを各 1 回用いたところ、次の脆弱性候補が報告された。モデル名は非公開で、深刻度も AI による分類である。[10]

モデル	Critical	High	Medium	Low	合計
Model 1	4	5	6	0	15
Model 2	3	5	4	1	13
Model 3	0	5	6	2	13

Model 1 には、似た問題を細分化して数えた影響があると説明されている。候補の内容、数え方、重大性の判断がモデルごとに異なるため、件数だけでは優劣を決められない。[10]

【実務への示唆】指摘候補の重複整理、再現確認、利用環境に照らした影響評価が必要となる。複数モデルが共通して指摘する問題は確認の優先候補になるが、一つのモデルだけが指摘した重大な問題を捨てる理由にはならない。知財調査でも、抽出件数や一致件数が多いことと、重要な見落としが少ないことは別であり、件数から検証できる内容へ評価を掘り下げる必要がある。

8. 英米の評価との比較と、公表までの時間差

英国 AISI は個別のサイバー課題に加え、模擬ネットワーク上で複数段階の作業を継続できるかも評価している。GLM-5.2 等について、同機関の評価方法では先行するクローズドモデルとの能力差が 4〜7 か月に縮まったと報告した。[11]

米国 NIST 掲載の 2026 年 9 月 17 日付評価では、後継の GLM-5.3 について、評価時点で最も高いサイバー能力を持つオープンウェイトモデルとしつつ、米国の最先端モデルとの間には依然差があるとしている。[12]

比較軸	日本 AISI の今回の評価	英国・米国の評価例
対象	V8 の既知脆弱性を使う能力。	複数の課題、模擬ネットワーク、複数ベンチマーク。
主な問い	どの技術的段階まで到達するか。	能力の広がり、長い作業の遂行、モデル間の差。
読み取れること	特定の高度な攻撃技術に残る限界。	より広いサイバー能力の進展。
時点	2026 年 6〜7 月の実験。	後継モデルを扱う 9 月公表の評価もある。

出典：[7][9][11][12]。異なる課題・実行条件の結果を、同じ尺度の順位として比較することはできない。

【評価】公表日の新しさと、実験結果の新しさを分ける必要がある。10 月 2 日に公表された日本のノートは 6〜7 月の評価であり、技術進展が速い分野では、実施から公表までの時間差自体が評価体制の改善課題になる。英米の評価と日本の評価の違いは、単純な矛盾ではなく、対象・条件・時点の違いを含む。

9. 企業が利用できるガイド・ツール・事業実証

AISI は、評価の観点、実際に評価する道具、組織として運用する仕組みを提供している。企業との接点が大きい成果物を整理する。

成果物・活動	内容と利用場面	文献
評価観点ガイド第 1.20 版	2026 年 7 月改訂。エージェントの「観測と制御」を追加。導入・更新時の評価項目の基礎。	[5]
AI セーフティ評価環境「aisev」	2025 年 9 月に OSS 公開。評価ツール・データセット、自動レッドチーミングを提供。	[13][14]
Chief AI Officer ガイド	2026 年 3 月公開。責任者、組織、調達、運用、監査等のガバナンスを扱う。	[15]
既知の攻撃と影響	2026 年 4 月に第 2 版。AI 特有の攻撃と影響を整理し、試験シナリオの作成を支援。	[6]
AI ロボティクスの評価観点ガイド	2026 年 7 月公開。人と空間を共有する AI ロボットの物理的・心理的・社会的リスクを扱う。	[16]

aisev は企業独自の評価への応用も想定される。ただし、公開リポジトリは、ツールの結果が IPA・AISI の公式認定・保証を意味しないと明示している。ツールの使用実績と対象システムの安全性の十分な検証は、別に説明する必要がある。[13][14]

2026 年 10 月 6 日に公開された事業実証 WG ビジョンペーパー第 2.0 版では、分野別のヘルスケア、ロボティクス、教育、ジオ AI と、分野横断のデータ品質、適合性評価の六つの SWG が示されている。教育とジオ AI が新たに加わった。[17][18]

2025 年度の活動は年次レポートとして総括されている。また、評価環境の改善については、2025 年 10 月から民間企業と共同で検討タスクフォースを組成し、2026 年 4 月に報

告書を公表した。ガイド・ツールの公開に加え、利用側の知見を反映させる活動が行われている。[23][24]

【評価】基盤モデルの性能試験だけでなく、個々の産業・利用場面で AI をどう評価するかにも重点が置かれている。医療文書作成、搬送ロボット、教育支援等では、同じ誤りでも影響と必要な人間の関与が異なる。分野別の実証と共通基盤を組み合わせる方向性は合理的と考える。

10. 国際連携と開発企業との協力

AISI は国際ネットワークでの共同テストに参加し、多言語評価や AI エージェント評価の知見を共有している。企業との協力では、Anthropic と 2025 年 10 月 29 日、OpenAI と 2026 年 5 月 29 日に協力覚書を締結した。[19][20][21]

協力先	公表された協力内容	確認範囲
Anthropic	評価手法、モデルの能力・限界・リスクに関する情報交換やベストプラクティスの共有。	協力覚書の公表概要。
OpenAI	安全性評価、ベンチマーク、サイバーセキュリティに関する協力可能性の検討・対話。	協力覚書の公表概要。
国際ネットワーク	多言語やエージェント等の共同テスト、評価手法・知見の共有。	共同テストの報告。

出典：[19][20][21]。公表概要だけでは、日本 AISI がすべての新モデルについて公開前アクセスや評価権を確保したとは確認できない。今回の研究条件で評価できたことと、恒常的な事前評価制度の成立は区別する必要がある。

11. 現状への評価と、今後確認すべき課題

【総合評価】実測評価への進展を評価しつつ、評価範囲と運用能力の拡充を確認すべき段階にある。以下は公開資料を踏まえた筆者の評価であり、AISI 自身による採点ではない。

評価する点	残る課題

評価する点	残る課題
モデル名と具体的な評価結果の公開。	対象モデル、分野、試験条件の拡大。
高いスコアの事例の再検証。	複数回試行、ばらつき、信頼区間の報告。
実行記録を通じた結果の解釈。	第三者が再検証できる設定情報・記録の公開範囲。
ガイド、ツール、業界別実証の組合せ。	企業での利用実績と改善効果を示す指標。
人員・予算の強化方針の具体化。	人数に加え、継続評価を担う専門能力の確保。
海外機関・開発企業との関係拡大。	協力関係と評価の独立性を両立する運営。

今後の評価報告では、次の五点を確認したい。

- モデルと設定の特定：正確なバージョン、推論設定、実行環境、安全制御の状態が明示されるか。
- 結果の安定性：何回試し、どの程度ばらつき、まれな高能力行動がどの頻度で現れるか。
- 実際の利用との対応：通常利用時の安全性と、安全制御を外した潜在能力をそれぞれ評価しているか。
- 試験対象の広さ：外部ツール、長時間作業、複数エージェント等を扱えるか。
- 公表・更新の速さ：新モデルや重大な挙動に対して迅速に追試・更新できるか。

これらは今回の結果を否定するものではなく、公的な評価機関の成果を意思決定に使えるものにするための条件である。

12. 企業の知財・研究開発部門への示唆

評価観点ガイド第 1.20 版は、アクセス権限の限定、ファイル操作の制限、実行履歴の観測、目標からの逸脱検知、人間による介入・停止等を評価項目に挙げる。[5]

以下は、これを知財業務に当てはめた筆者の実務提案である。AISI が知財業務専用の要件として定めたものではない。

知財業務での利用	評価すべき事項の例
特許 SDI・競合監視	検索・分類の再現性、重要案件の見落とし、根拠公報への追跡可能性。
発明提案・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。
FTO・クレーム対比	限定事項の脱落、法的状態の取り違え、不確実な判定の明示。
社内外の情報を横断するエージェント	閲覧権限、外部文書に埋め込まれた指示への耐性、情報持出し。
出願・案件管理の自動化	書込み権限、重要操作の承認、誤操作からの復旧、監査記録。

導入判断では、どのモデルを使うかに加えて、どの情報を読ませ、どのツールを使わせ、どこまで実行させるかを評価する必要がある。ローカルで動かす場合も、データの外部送信を抑える効果と、誤動作・権限逸脱を防ぐ設計は別に検証することになる。

【知財戦略への示唆】AI 安全性評価のデータ、評価シナリオ、制御技術、監査・運用ノウハウは、企業の競争力につながる可能性がある。AISI の適合性評価 SWG は国際標準化動向も扱っており、評価方法が調達条件や業界の共通要件へ発展するかは注視すべきである。ただし、具体的な認証義務や調達要件がすでに決まったという意味ではない。[22]

13. 今後 1〜2 年の見通し

事業実証ロードマップは、2026〜2027 年度に評価手法の精緻化、評価環境の整備、AI エージェントへの対応、国際標準との整合を進める方向を示している。[18]

以下は公開された方向性を踏まえた予測であり、確定した制度や日程ではない。

展開	見通し
個別モデルの評価公表の継続	可能性が高い。更新の速さと対象の広がりが焦点となる。
エージェント全体の評価	モデルとツール、権限、記憶を組み合わせた検証の重要性が増す。
業界別の評価データ・試験環境	事業実証 WG を通じて具体化が進むと予想する。

展開	見通し
企業の AI 調達・監査への活用	ガイドや評価結果が要求仕様や社内審査の参考になる可能性がある。
認証・相互承認の枠組み	検討は進むが、実施時期・法的効果は現段階で断定できない。

AISI の今後の成果は、公表文書数や職員数に加え、評価結果が企業の設計変更、事故予防、導入判断にどの程度使われたかで判断するのが適切である。

調査対象と確認範囲の補足

依頼で示された動画は「日本版 AISI が本格始動…Claude の『攻撃能力』を検証した結果」（https://www.youtube.com/watch?v=y-aWZ820gGM）である。動画の音声・字幕全文は未取得のため、その全発言の検証や逐語引用は行っていない。対応する解説記事は確認した。[25]

Japan AISI 公式 X および塩崎あきひさ氏の 2026 年 10 月 5 日投稿は、依頼文の URL が省略されていた。検索結果と関連記事を補助的に確認したが、投稿全文の取得に制約があった。このため、組織の現状や技術的評価の根拠を当該投稿だけに依存させていない。

最新の職員実数、予算執行額、非公開モデルの恒常的な事前評価権、非公開の試験結果については、本調査で確定していない。公表時点と評価実施時点、方針と実績、モデルの潜在能力と通常サービスの挙動を区別した。

参考文献

すべてのウェブ資料の最終確認日：2026 年 10 月 6 日。公開日と評価実施日は区別して記載した。本文中の［番号］は以下の文献に対応する。

[1] AI セーフティ・インスティテュート．AISI について．公開日記載なし．

設立経緯、所長、業務。

https://aisi.go.jp/about/

[2] AI セーフティ・インスティテュート事務局．AI セーフティ・インスティテュート（AISI）について．2026 年 7 月 1 日（資料表記）．

18 頁。特に組織・関係府省庁・推進体制。

https://aisi.go.jp/assets/pdf/20260708_AISI_jp.pdf

[3] 内閣府 人工知能政策推進室．AISI 機能強化の方向性（案）．2026 年 4 月 7 日．

AI 戦略推進会議幹事会第 3 回・資料 1-1、28 頁。特に 3 頁（人員・予算比較）、7〜8 頁（機能強化・出向）、28 頁（当時の評価）。

https://www8.cao.go.jp/cstp/ai/aisi/3kai/renrakukaigi_manager/shiryo1_1.pdf

[4] AI セーフティ・インスティテュート．AISI 事業実証 WG 下期報告会・今後の方向性に関する資料．2026 年 3 月 10 日．

13 頁。自ら評価する能力の獲得、官民共創による評価基盤の方向性。

https://aisi.go.jp/assets/pdf/0311_08_future_direction.pdf

[5] AI セーフティ・インスティテュート．AI セーフティに関する評価観点ガイド（第 1.20 版）．2026 年 7 月 7 日．

66 頁。特に 3.11「観測と制御」、自律的な挙動、外部環境との相互作用。

https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.20_ja.pdf

[6] AI セーフティ・インスティテュート．AI システムに対する既知の攻撃と影響．2026 年 4 月 24 日（第 2 版への更新）．

AI システム特有の攻撃と影響、レッドチーミングへの利用。

https://aisi.go.jp/output/output_security/known_attacks_and_impacts/

[7] AI セーフティ・インスティテュート．AI 検証ノート：ExploitBench による Claude Opus 4.8 のサイバー能力評価 〜事例分析の重要性について〜．2026 年 10 月 2 日．

評価実施は 2026 年 6 月。評価設定、前モデル比較、pc_control の事例、追試、免責事項。

https://aisi.go.jp/activity/activity_notes/261002_1/

[8] Seunghyun Lee; David Brumley. ExploitBench: A Capability Ladder Benchmark for LLM Cybersecurity Agents. 2026 年 5 月 13 日（v1）.

arXiv:2605.14153。DOI: 10.48550/arXiv.2605.14153。プレプリント。本文：

https://arxiv.org/html/2605.14153v1 。

https://arxiv.org/abs/2605.14153

[9] AI セーフティ・インスティテュート. AI 検証ノート：高度な AI サイバー能力のオープンモデルへの拡がり. 2026 年 10 月 2 日.

評価実施は 2026 年 7 月。GLM-5.2 と Opus 系モデルの比較、ナッジ、安全上の拒否応答。

https://aisi.go.jp/activity/activity_notes/261002_2/

[10] AI セーフティ・インスティテュート. AI 検証ノート：AI が見つける脆弱性に違い！. 2026 年 7 月 8 日.

匿名 3 モデルによる試行的な脆弱性検出比較。

https://aisi.go.jp/activity/activity_notes/260708/

[11] UK AI Security Institute. How Far Behind the Frontier are Leading Open Weight Models on Cyber?. 公開日の確定情報は今回未取得.

GLM-5.2、DeepSeek V4-Pro 等の個別課題・模擬ネットワーク評価と限界。

https://www.aisi.gov.uk/blog/how-far-behind-the-frontier-are-leading-open-weight-models-on-cyber

[12] National Institute of Standards and Technology（NIST）／CAISI. CAISI's Assessment of Z.ai's GLM-5.3 Cyber Capabilities. 2026 年 9 月 17 日.

4 ベンチマークによるサイバー能力評価。機関名は記事公表時の表記。

https://www.nist.gov/news-events/news/2026/09/caisis-assessment-zais-glm-53-cyber-capabilities

[13] AI セーフティ・インスティテュート. AI セーフティ評価のための評価ツールを OSS として公開. 2025 年 9 月 12 日.

評価環境、自動レッドチーミング、Apache 2.0 による公開。

https://aisi.go.jp/output/output_information/250912/

[14] Japan-AISI. aisev：AI セーフティ評価環境（GitHub リポジトリ）. 更新型資料.

README、評価環境の概要、試作版・認定保証に関する注意。

https://github.com/Japan-AISI/aisev

[15] AI セーフティ・インスティテュート. Chief AI Officer ガイド. 2026 年 3 月 17 日.

ガイドブックおよび CAIO 設置・AI ガバナンス実務マニュアルの公開ページ。

https://aisi.go.jp/output/output_framework/260317/

[16] AI セーフティ・インスティテュート．AI ロボティクスに関するセーフティ評価観点ガイドの公開．2026 年 7 月 23 日．

人と空間を共有する AI ロボットの評価観点。

https://aisi.go.jp/output/output_information/260723/

[17] AI セーフティ・インスティテュート．AISI 事業実証ワーキンググループ・ビジョンペーパー（第 2.0 版）～信頼とイノベーションが両立する AI 社会の実現に向けて～ の公開．2026 年 10 月 6 日．

教育・ジオ AI を含む新たな事業実証体制。

https://aisi.go.jp/output/output_information/261006/

[18] AI セーフティ・インスティテュート 事業実証ワーキンググループ．ビジョンペーパー（第 2.0 版）概要版．2026 年 10 月 6 日公表．

17 頁。六つの SWG、2026～2027 年度を含むロードマップ。

https://aisi.go.jp/assets/pdf/VisionPaper_Summary_v2.0_jp.pdf

[19] AI セーフティ・インスティテュート．共同テスト演習の評価レポートが公表されました。．2025 年 7 月 18 日．

国際ネットワークの多言語評価・エージェント評価。

https://aisi.go.jp/activity/activity_international/250718/

[20] AI セーフティ・インスティテュート．米 Anthropic 社と MOC を締結しました。．2025 年 10 月 29 日．

協力覚書の公表概要。

https://aisi.go.jp/activity/activity_international/251029/

[21] AI セーフティ・インスティテュート．米 OpenAI 社と MOC を締結しました。．2026 年 5 月 29 日．

安全性評価、ベンチマーク、サイバーセキュリティ等の協力に関する公表概要。

https://aisi.go.jp/activity/activity_international/260529/

[22] AI セーフティ・インスティテュート．2025 年度 適合性評価 SWG 活動報告書の公開．2026 年 4 月 23 日．

適合性評価制度、国際標準化、Human-Machine Teaming に関する検討。

https://aisi.go.jp/output/output_information/260423/

[23] AI セーフティ・インスティテュート．AI セーフティ年次レポート 2025．2026 年 4 月 28 日．

2025 年度活動の総括とファクトシートの公開ページ。

https://aisi.go.jp/output/output_information/260428/

[24] AI セーフティ・インスティテュート．「AI セーフティ評価環境 検討タスクフォース」の「検討報告書（2025 年度版）」の公開．2026 年 4 月 15 日．

評価ツールの改善に向けた企業との共同検討。

https://aisi.go.jp/output/output_information/260415_2/

[25] AI 時短ラボ．日本版 AISI が本格始動、初めてモデル名を出して Claude Opus 4.8 と GLM-5.2 のサイバー能力を評価．2026 年 10 月 6 日．

ご提示の動画に対応する解説記事。動画・SNS の背景確認に補助的に使用。技術的結論は一次資料を優先。

https://www.ai-jitan-hub.com/news/japan-aisi-claude-opus-4-8-glm-5-2-cyber-evaluation