資料２

AISIにおける
サイバーセキュリティに資する取組

令和8年３月

内閣府人工知能政策推進室

AIセーフティ・インスティテュート（AISI）について

- AIが社会に与える影響が拡大するなかで、その**安全性と信頼性を専門的かつ中立的な立場で検証する「公的な第三者機関」であるAIセーフティ・インスティテュート（AISI）が必要不可欠に**。国際的にAIガバナンスの重要性が高まる中、**AI安全性サミット**（2023年11月、英国）を契機に**「AI安全性」を具現化するための議論**が進み、英・米はAISIを設置。
- 我が国も第7回AI戦略会議（2023年12月）における**岸田元総理からの指示を踏まえ**、2024年2月に日本の**AIセーフティ・インスティテュート（所長：村上明子氏）を設置**。
- AI安全性の知見のハブとして、**国内外の関係機関とのネットワークを強化中。AI安全性の評価能力を確立**しながら、**AI安全性評価のための基準、ガイダンスを作成**。

日本のAISI（エイシー）の概要

名　称	（日本語）AIセーフティ・インスティテュート （英　語）Japan AI Safety Institute（略称 J-AISI）
業務内容	● 安全性評価に係る調査、基準等の検討 ● 安全性評価の実施手法に関する検討 ● 他国の関係機関（英米のAI Safety Institute等）との国際連携に関する業務
関係機関	内閣府、国家安全保障局、内閣サイバーセキュリティセンター、警察庁、デジタル庁、総務省、外務省、文科省、厚労省、農水省、経産省、国交省、防衛省 情報通信研究機構、理化学研究所、国立情報学研究所、産業技術総合研究所、情報処理推進機構
主要な実績	**AIセーフティに関する評価観点ガイド、レッドチーミング手法ガイド等の公開。AISI国際ネットワーク**(米・EU等の主要国AISI関連機関10カ国が参加)**に参加**し、AIの共同テストに関するトラックをリード。

村上 明子
(AIセーフティ・インスティテュート所長)
(SOMPOホールディングス株式会社執行役員常務　グループChief　Data Officer)

2024年、AIの安全性と信頼性を専門的かつ中立な立場で検証する公的な第三者機関であるAIセーフティ・インスティテュート(AISI)の設立に伴い所長に就任。

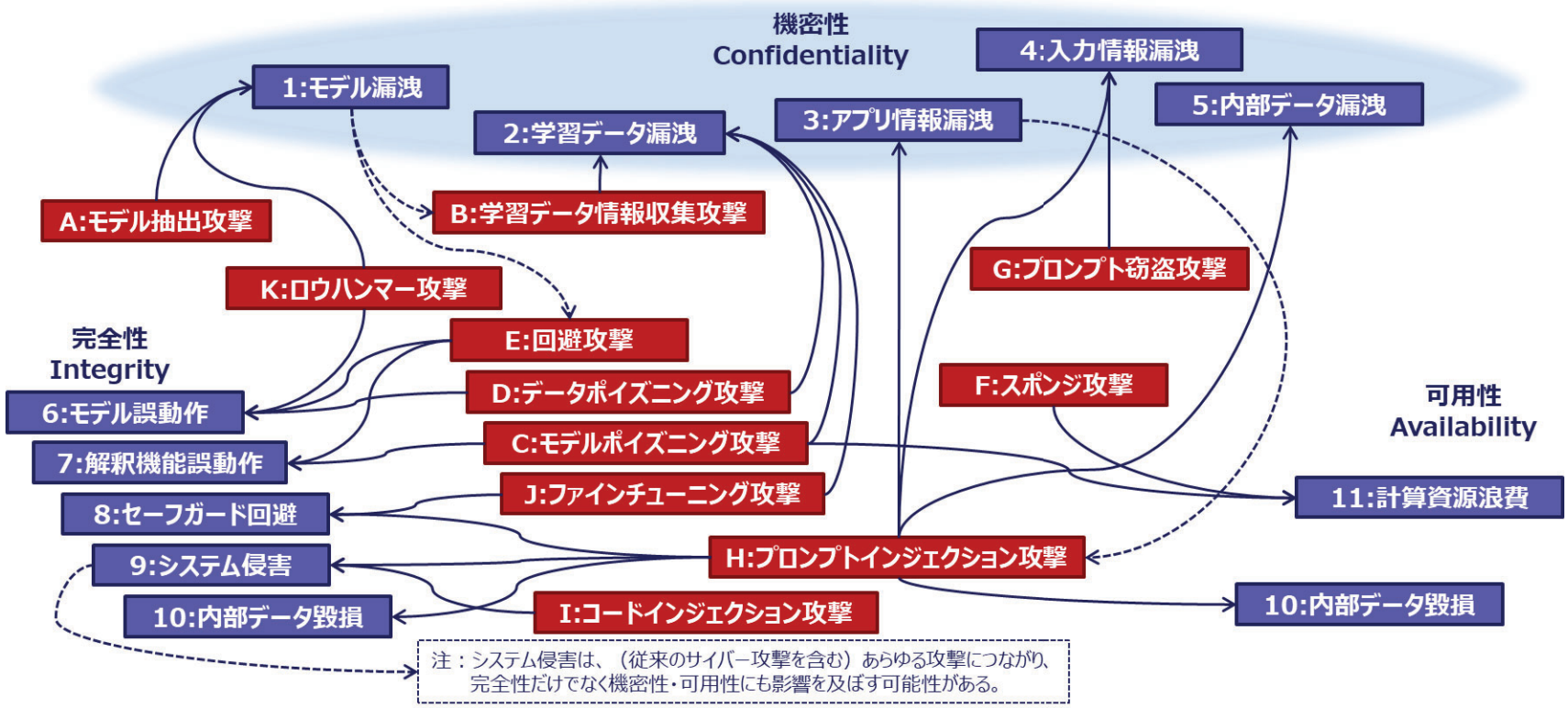
これまでのAISIにおけるAIセキュリティに係る主な取組

「AIシステムに対する既知の攻撃と影響」（2025年３月）

ＡＩシステムに対する特有のセキュリティ攻撃を俯瞰すべく、**学術論文等で発表されたAIやAIシステムに対する攻撃とその影響を取りまとめ**

「AIインシデントレスポンス・アプローチブック」（2026年1月9日公表）

ＡＩインシデント発生時の被害を最小化することを目的に、**情報システム向けインシデントレスポンス（NIST SP800-61）を土台に、ＡＩ特有の対策を追加**したものを取りまとめ



サイバー攻撃とAIへの影響の関係

主要国AISI関連機関の動向：「セキュリティ」や「標準」に重点

- 2025年2月14日、**英国AISI**はAI Security Institute に改名。**AI安全性(AIモデルの透明性、堅牢性、信頼性等)に係る取組からセキュリティ(サイバー・化学攻撃等への活用の可能性、犯罪への活用の評価等）にシフト。**防衛科学技術研究所とも連携。
- **2026年2月26日、詐欺及びサイバー犯罪によるAIの誤用に係る評価フレームワークを公表**。14LLMについて2万件以上の評価を行い、不正利用支援に最も影響を与える要因を特定。

- 2025年6月3日、**米国AISIは**CAISI（Center for AI Standards and Innovation） に改名。**AI安全性に係る取組から、よりイノベーション促進と国家安全保障に重点をおいた取組（①セキュリティ評価機能の強化・拡充**（敵対国のAIシステム評価、バックドア対策等）、**②外国の脅威への対処**（競争力評価、外国製AIシステム採用状況調査等）**にシフト**。
- 同年9月30日、**中国DeepSeekの欠点とリスクを指摘する評価を公表**。セキュリティの欠陥や検閲は、開発者や消費者、米国の国家安全保障にリスクをもたらす可能性があると指摘。

- 2025年7月26日、**中国**は、上海での世界人工知能大会で、**中国主導で、ＡＩを安全かつ包摂的に活用するための国際組織の創設を提唱**。
- 2026年1月に**改正サイバーセキュリティ法を施行し、AI技術の開発と安全を監視するための法的基盤を確立。**AIの基礎理論、アルゴリズム研究及びインフラ構築の支援を明文化する一方で、倫理規範、リスク監視、および安全監督の強化を義務付け。

- 2025年7月30日、**英国AISI及び加AISIは、30億円規模の基金を設け、急速な進化を遂げるAIが、人間の価値観や倫理観に沿って行動するよう管理・制御するための研究開発**を、多国間の官民連携で実施する「Alignment Project」の開始を表明。

※**2025年に12月に英国AISIは世界のフロンティアAIモデルの性能評価**（安全性・セキュリティ面ほか）を、**米国AISIはサイバー攻撃等に対するAIシステムの堅牢性の計測手法**の開発など実施。

日本AISIと米国アンソロピック社の協力について

1.令和7年10月29日に、アンソロピック社のダリア・アモデイCEOは、高市総理大臣を表敬訪問。小野田内閣府特命担当大臣（人工知能戦略）が同席し、MoCに基づきAISI村上所長も同席。

高市総理は以下について発言。

「日本における信頼できるＡＩの実現に向けて、安全に関するＡＩＳＩ（エイシー）との協力、政府での活用、スタートアップ支援で協力いただけることに感謝。是非、日本での開発拠点等更なる投資に期待したい。」

2.同日、**AISIと米国アンソロピック社はAIの評価に関する協力覚書(MoC)を締結し**、以下について協力。

（主な協力内容）

- AIモデル評価に関する情報やベストプラクティスの共有
- AIモデルの能力やリスクを評価するためのツールやベンチマークの開発
- AI分野の動向や将来の技術開発に関する意見交換

アンソロピック　脅威インテリジェンス・ブリーフィング
(2025.11「初めて報告されたAI主導型サイバー諜報活動の阻止」)

1. **Claude**（アンソロピック社の生成AI)**を活用したサイバー攻撃が複数報告**。攻撃手法が**驚異的なスピードで大きく進化**。人間の関与が10~20%に留まり、**AIが自律的にサイバー攻撃するフェーズへ**。**技術や資金の少ない攻撃者でも、大規模かつ効率的なサイバー攻撃が可能**に。

2025年3月

- 英国拠点の脅威アクターがClaudeを活用し、**技術力不足を補い、ノーコードでランサムウェアを開発**。
- ダークウェブで高度なマルウェアを流通・販売（$400~1,200)。

2025年5月

- ロシア語を話すサイバー犯罪者がClaude Code（アンソロピック社のAIエージェント型コーディング支援ツール）を使い、国内外の17の標的に対して**大規模な恐喝を実施**。（要求額：ビットコインで＄75,000~500,000)
- Claude Codeが大規模な偵察、認証情報等の収集、ネットワーク侵入を自動化。

2025年9月

- 中国政府支援グループがClaude Codeを使い、**自律型サイバー攻撃エージェントを構築**。Claudeをオーケストレーションシステムとして用い、複雑な多段階攻撃を個別の技術タスク（脆弱性スキャン、認証情報の検証、データ抽出、横展開等）に分解することで、悪用検知が非常に困難に。
- **サイバーキルチェーン全体（脆弱性発見、侵入、自律的分析、横展開、権限昇格、情報流出）を概ねAIが自律的に実行**。

2. **AIは防御にも不可欠。AIを使った高度な侵入検知技術の向上やAIを使った自動診断・自動パッチシステムなどの安全対策の強化**がますます重要に。

出典：Anthropic, “Disrupting the first reported AI-orchestrated cyber espionage campaign”, https://www.anthropic.com/news/disrupting-AI-espior

日本AISI機能強化の必要性

■ 2023年11月、英国でのAI安全性サミットを契機に、英・米がそれぞれ国内にAISIを設立。
日本も、**2024年2月に日本AISIをIPAに設置。AISI国際ネットワークを形成。**

■ **日本AISIの予算、人員は英米に比べて圧倒的に少ない。**
国際ルール形成主導に向け、**産官学の人材、知見、資金を糾合して機能を強化する必要。**

	日本 AI Safety Institute	英国 AI Security Institute	米国 Center for AI Standards and Innovation
設立	2024年2月	2023年11月	2024年2月（25年6月にCAISIへ改名）
所管	経済産業省・デジ庁 IPA（情報処理推進機構）内	科学・イノベーション・技術省	商務省（NIST内）
所長	村上明子	Adam Beaumont	Austin Mayron
職員数・予算	**31人（併任含む*）** 令和6年補正：**3.8億円** ＊IPAや理研からの併任	**約200名人以上*(うち専門家は90人)(目標300人)** 初期予算：**£1億（約200億円）** *2025年9月時点、大学教授、元Google, 元Open AI等トップ人材を採用 *トップAIモデルへの特権アクセス及びコンピューティングへの優先的アクセス	**30人程度（目標80人）**** 2024年度:**予算＄1000万(約15億円)** **2024年時点情報
役割	• AI事業者ガイドラインの策定支援、米国ガイドラインとの相互比較を実施。 • 評価観点ガイド、レッドチーミング手法ガイドなど実務ドキュメント作成。	• フロンティアモデルの評価ベンチマークとテストプラットフォーム構築。 • AI安全性・セキュリティの最新研究の白書の発行。 • 米国AISIとの共同テスト、カナダAISIとの協力などAISIネットワークのハブ。	• OpenAI・Anthropic等との間で、フロンティアモデルの事前評価（プレリリース・テスト）協定。 • モデル評価・リスク管理の技術スタンダードを策定。 • Google、Microsoft、Anthropic 等200社超を巻き込んだ共同研究。

令和7年度補正予算を活用したAISIの機能強化

AISIの機能強化を加速するため、令和7年度補正予算合計88億円を措置

AISIが自ら評価ツールを開発。我が国で活用されるAIについて、**セーフティのみならず、セキュリティの観点を含めて分析・評価する能力を持つ。**
(AI評価手法の構築、専用テストベッド・計算資源の整備等)

【具体策】

- **日本語に関する出力データの安全性確認（AIセキュリティ含む）**を中心に、**AI評価手法を開発**し、民間企業等に提供
- **AIエージェントに係る安全評価ガイドライン**や、適正性を評価するためのチェックツールを開発・提供。
- **産総研に委託し、人との協働ロボット等の安全性に係る研究開発、ガイドラインの策定、更に国際標準化に向けた検討を**実施。
- **国内外の民間企業等が開発したフロンティアAIやAIエージェント、フィジカルAIの安全性・セキュリティを事前に評価**するための調査**(専用テストベッドの整備に向けた取組)**

今後のAISIの機能強化の方向性

ＡＩ基本計画や総理指示、自民党からのAISIの提言を踏まえつつ、**人員の増強や国研等のパートナーシップ機関との連携強化**を図りながら、**AISIにおいてＡＩモデルの技術的評価、広範な適正性に係る評価、セキュリティ面での対策を実行できる体制を構築**

(「ＡＩ基本計画 (令和7年12月閣議決定）」（抄）)

ＡＩイノベーションの好循環を実現し、信頼できるＡＩエコシステムを構築するため、技術開発・実証・評価・運用の各段階において、適正性の確保につながるＰＤＣＡサイクルを 構築する。

これを実現するため、国民や事業者等の自主的かつ能動的な取組を促すよう、国としての基本的な考え方を提示する。当該考え方等を踏まえ、ＡＩセーフティ・インスティテュート （ＡＩＳＩ）を抜本的に強化することで、ＡＩモデルの技術的評価を適切に行い、当該評価も踏まえ、ＡＩがもたらすリスクに係る実態把握を行うとともに必要な措置を講ずる。ＡＩＳＩの機能強化にあっては、世界屈指の英国 AI Security Institute の規模 をベンチマークとしつつ、人員を直ちに現行の２倍程度に拡充する。

ＡＩの安全性確保やＡＩを利用した攻撃への対応が、新たなサイバーセキュリティ上の課題として認識されつつあることを踏まえ、体制整備を含めた適切な措置を講ずる。

（１） 信頼できるＡＩエコシステムの構築

ＡＩモデルの安全性にとどまらず、より広範な適正性に係る評価やセキュリティ面での対策を実行できる体制を構築し、技術的・制度的なガバナンスの強化を図る。 その中核として、ＡＩＳＩの機能を、政府を挙げて抜本的に強化する。

（２） ＡＳＥＡＮ等グローバルサウス諸国を含めた国際協調

広島ＡＩプロセスの推進や、ＡＩＳＩネットワーク等の国際的な枠組みの活用により、ＡＩガバナンスの構築を主導する。