

AISI事業実証WGへの期待



令和8年10月
内閣府人工知能政策推進室
齊藤大地



最新のトレンド（AIリスク・脅威論争の歩み）

2019

GPT-2 リリース論争

当時OpenAIの研究チームに在籍していたDario Amodei氏主導で、開発されたGPT-2について、**OpenAI社内で、危険すぎるとして技術を公開すべきではないとする派閥と、技術を公開すべきとする派閥の間で対立が発生**。Amodei氏は慎重派。

2026.4

Claude Mythos登場

Anthropic社の最新鋭モデルが、サイバー性能評価で突出した成績を記録し、**高性能AIによるサイバー攻撃の懸念が、現実的な危険として顕在化**。

また、**生物・化学テロ等、他のドメインでの悪用の懸念も増加**。

2026.9

RSI※論争勃発

※再帰的自己能力改善

（Recursive Self-Improvement）

Anthropic社の研究者、Jacob Coxon氏が高性能AIの**破滅的危険**について警鐘を鳴らし、辞任。

米産業界・政界を中心にAIの危険性、特に**RSI（再帰的自己能力改善）**について関心が高まり、規制論についても賛成派、反対派で論争が勃発。

コーディングエージェント登場

Anthropic の Claude 3.5 SonnetやOpenAI の o1AI等、コード生成が可能なモデルが出始め、次第にチャットボットからAIEージェントへ。

2025年にAnthropicのClaude CodeやOpenAIのCodex Agent等の「ハネス」ソフトウェアが普及し始め、エージェント能力が飛躍的に向上。

2024

OpenAI-HuggingFace事案

OpenAI社内のモデル性能試験中に、隔離環境から脱出したAIモデルが、別企業（HuggingFace社）のサーバーに不正に侵入する事案が発生。

後の調査報告書で、AIEージェントの群れがOpenAI社のサーバー内で情報共有を行う「掲示板」を設置し、謀議していたこと等が判明し、**自律型AIEージェントの暴走によるリスク**に大きな関心が集まる。

2026.7

(参考) 高性能AIの自律的サイバー攻撃が現実に (2026年7月)

- 昨年段階のAIが人間によるサイバー攻撃を支援する、今春のA I エージェントが人間の目標設定を踏まえて攻撃工程を自動化する、から踏み込んで、**現在、人間が明示しない行動まで選び、自律的に攻撃する状態へ**

Anthropic
Opus4.7/Mythos5
/内部研究モデル

7月30日公表

実世界への越境

サイバー試験中のAIが、実験用ではない実在組織へアクセス

- 企業のデータを閲覧
- 悪意あるプログラムを公開
- 約9,000のシステムを探索

OpenAI

7月9～13日

実世界への侵入

AIが実験環境の未知の弱点を自ら見つけ、外部へ出てHugging Faceに侵入

- 約17,600の操作を確認
- パスワード等を取得し、別システムへ移動
- GLM-5.2を事後解析に活用

Kimi K3
(米CAISIと英AISIの調査)

7月23日公表

能力評価

企業ネットワークを模した試験で攻撃能力を測定。

- 32段階中、平均17段階まで到達
- 10回中1回は最後まで成功
- オープンウェイトの攻撃能力向上を確認

OpenAIGPT-5.6 Sol
AnthropicのMythos 5
(英AISIの調査)

8月4日公表

想定外行動の評価

AIが指示されていない外部行動を選ぶ事例を確認。

- 122回中10回で発生
- 計19件の許可されていない行動
- 偽人物像、情報を盗むメール等

※内閣府AI室においてもAI法16条に基づく個別ヒアリングを実施

- 9月24日、豪州政府より、OpenAI社のAIモデルが豪州政府のシステム上の非公開のデータにアクセスしたとの発表（詳細は随時公表されている状況で全体像不明）。

■ 本大統領令の「フロンティアAIの**「自主的な(評価)枠組み」**の施行が8月初旬から始まる。

1. 高度なAIに向けた米国システムのアップデート

- 本大統領令発令から**30日以内**に、
 - （1）国家安全保障システム委員会は、**国家安全保障システムのサイバー防衛**を優先的に実施。
 - （2）戦争長官は、**戦争省の情報システムのサイバー防衛**を優先的に実施。
 - （3）国土安全保障長官は、**連邦政府の情報システムのサイバー防衛等に関する拘束力のある適切な運用指針**を公表。
 - （4）財務長官及び国土安全保障長官は、**AI産業及び重要インフラの運営事業者と自主的に連携し、ソフトウェア脆弱性の検証・修正、脆弱性パッチの配布の調整等を行う「AIサイバーセキュリティ・クリアリングハウス」を設立。**

2. セキュアなフロンティアモデルの展開

- 本大統領令発令から**60日以内**に、財務長官、戦争長官及び国土安全保障長官は、
 - （1）AIモデルの高度なサイバー能力を評価し、**特定のフロンティアモデルの閾値を決定するベンチマークプロセス**（非開示）を策定。
 - （2）**AI開発者と協力し、開発者が以下を行うことができる自主的な枠組みを策定。**
 - ア 開発中のモデルが**特定のフロンティアモデル**の閾値を満たすか決定するため、**連邦政府と協議。**
 - イ 特定のフロンティアモデルへのアクセスを、他の信頼できるパートナーに公開する前に**最大30日間、連邦政府に提供**
 - ウ 連邦政府と連携し、特定のフロンティアモデルに**早期アクセス可能な信頼できるパートナーを選定**
- 注：本節のいかなる規定も、フロンティアモデルを含む新しいAIモデルの開発・公表等に関する政府による**強制的なライセンス、事前承認又は許可要件を設けることを認めるものと解してはならない。**

- 当初米国の報道では、**「オープンウェイトは対象外」とされる。**「クローズドソースで、最先端能力を持ち、国家安全保障リスクを伴うモデルに限定している」と関係者が説明している由。「（非開示の）ベンチマーク」でオープンウェイトを評価しないかは不明。
- NISTでは、出自に関わらずあらゆる国・地域から登場するオープンウェイトモデルについて追跡・監視。モデル評価には、公開・非公開両方のベンチマークを活用。

最新のトレンド（オープンセキュアAIアライアンス（NVIDIA））

- 26年7月27日、**NVIDIAが呼びかけ、120超の企業、団体が立ち上げから参加**（Microsoft、IBM、Hugging Face、Linux財団、Cisco、Palantir等。日本からはNTTのみが参加）
- **オープンウェイトモデルAIを使ってサイバー防御そのものを強くするための産業連合（AIモデルのユーザーを含む）**で、オープンな防御技術・評価手法・エージェント基盤を共同開発。
- Hugging Face事件が影響、「**オープンなモデルなら、防御側が中身を検証し、自社環境に適応し、自ら運用できるため、少数のclosed modelだけに依存するよりも防御能力を分散できる**」と主張。
- AIEージェントの安全をモデル単体ではなく、①ID・認証、②権限管理、③ハーネス、④ガードレール、⑤ログ、⑥評価、⑦安全なモデル形式、⑧脆弱性検査・安全なコーディングまで、「**Agent Stack全体（下記）で確保**」（NVIDIAがオープンウェイトモデル、データ、エージェントハーネスの研究成果を提供）
- 日本を含めて**各国政府・企業が連携に関心（日本もAISIとの連携を探求中）**

AI Agent を構成する要素（Agent Stack）



最新のトレンド（S Iに関するフロンティア責任に関する共同コミットメント）

- トランプ大統領は2026年9月29日に、ホワイトハウスで**主要なAI・テック企業6社※**のCEOらが「**超知能に関するホワイトハウス協定：フロンティア責任に関する共同コミットメント**（White House Accord on Super Intelligence: Joint Commitment on Frontier Responsibilities）」を**公表し、署名**。
※グーグル、アンソロピック、メタ、オープンAI、xAI、エヌビディア
- 業界による自主的な協定であり、**新たなグローバル規制機関を設けることなく、業界の自主規制と多層的なレビューによって、米国が超知能（Super Intelligence）の開発競争で先行し続けることが目的**。
- トランプ大統領は「道義的拘束力があり、ほとんど憲法のようなものだ」「**10人程度から成る委員会**のようなものを設けることも検討」と説明。

【協定概要】

○フロンティアモデルを訓練・展開するすべての企業は、次の4層の管理体制を整備すべき。

1. 訓練時および展開時の社内統制

能力とアラインメントを監視すること。特に、サイバーセキュリティ、バイオセキュリティ、化学的脅威に重点を置くこと。モデルがハッキングを行ったり、意図しない形でシステムにアクセスしたりしないように止めること。

2. 社内の専任チーム

上記の統制が実際に機能していることを確認し、問題があれば是正する責任を持つチームを設けること。

3. 独立した外部監査人・評価者

統制や検知の仕組みが、意図されたとおりに機能しているかを外部から検証すること。

4. 独立した取締役会委員会

社内チームおよび外部監査人から報告を受け、問題が確実に是正されるよう監督すること。

○参加企業は、自らのシステムの**安全性向上のための基準及びベストプラクティスを策定**するため、定期的に会合を開催。

○**将来的には、これらの措置を法律又は規制として制度化することが適当となる可能性がある**。法的義務の有無にかかわらず、各社はその実施にコミットする。

第Ⅱ期人工知能基本計画（概要）

～日本A X、より強く、より豊かに～

令和8年7月14日 閣議決定

基本構想

- ◎急伸するエージェント型A Iに適応し、計算資源や電力から、制度・政策まで日本のA I実装能力を抜本強化
- ◎A Iを前提として意思決定や業務の進め方を根底から見直す「A Iトランスフォーメーション」（A X）を推進
- ◎人が人として創造すべき価値を探求し、「人間力」を増進しながら、「人とA Iが協働する社会」を率先して実現
- ◎「信頼できるA I」で社会全体を駆動し、課題解決と国力強化を探求、日本独自の価値を創出

4つの原則

- イノベーション促進とリスク対応の両立
- 挑戦と学習
- アジャイル（柔軟かつ迅速）な対応
- 内外一体での政策推進

4つの基本的な方針に基づく施策

政府一丸でA X推進に必要な体制を構築、あらゆる制度や運用を先導見直し

1. A I利活用の加速的推進「A Iを使う」：エージェント型A I等の積極適用、政府における率先導入

- 政府・自治体でのA Iの徹底した利活用 ガバメントA I「源内」の進化と普及、地方の伸び代を成長に転換（地域A X）
- A I利活用促進による社会課題解決、新事業創出 バーティカルA I・フィジカルA I中心に社会実装加速
- データ戦略、更なるA I利活用に向けた仕組みづくり 政府・準公共・産業分野でデータ連携基盤構築

2. A I開発力の戦略的強化「A Iを創る」：A I実装能力を高め、「A I主権」を確保する

- A I研究開発・利用基盤の拡充・高度化 国家基盤として計算資源やデータ、安定した電力供給を確保
- 日本国内のA I開発力の強化 データセット創出、内外トップ人材確保、モデル開発高度化、評価基盤・テストベッド整備
- 信頼できるA I基盤モデル等の開発 日本の文化等を踏まえたA I開発・評価、マルチモーダル基盤モデル開発推進
- 日本の勝ち筋の追求 バーティカルA I（領域別戦略）、フィジカルA I（A Iロボティクス、自動運転）、AI for Science

利活用
と
技術革新
の
好循環

3. A Iガバナンスの主導「A Iの信頼性を高める」：A Iエコシステムにおける「信頼」の構築で世界をリード

- 責任あるアジャイル・ガバナンスの実現 AI法を含めた制度等の能動的かつ不断の見直し、Project YATA-Shield
- A S E A N等グローバルサウス諸国を含めた国際協調 AISI機能強化、信頼できるA I共創に向けA Iサミットの早期日本開催

4. A I社会に向けた継続的変革「A Iと協働する」：人が意思決定への責任を持ちながら、A Xで日本を駆動

- A Iを前提とした産業構造・雇用環境の構築 中小企業、地方含むA X推進（取組可視化等）、雇用への影響を調査研究
- A I社会における制度・枠組みの検討・実証 規制・制度改革、人とA Iの責任関係、知財保護・利活用につながる透明性確保
- A I時代の人材の育成・確保 A I実装人材育成・確保、全ての国民のA Iに関するリテラシー向上、リ・スキリング支援
- A I時代における人間力の向上 創造力、批判的思考、判断責任など人間力向上、人とA Iの役割分担の探究

➡ 本基本計画は当面毎年変更、適切なベンチマーク（KPI等）を設定しモニタリング

バーティカルAIの官民投資ロードマップの方向性

現状認識、日本の強み

- バーティカルAIはデータ、モデル、アプリを垂直統合した領域特化型システム。高い正確性と専門性から「現場で使えるAI」。
注：2024年の国内バーティカルAI市場は約1兆円、2030年には約3兆円と見通し。
- AIエージェントの浸透や**AIを軸とした経営改革（AX）への展開**により、想定以上の付加価値の創出が期待されている。**フィジカルAIに先行して実装が進むバーティカルAIの現場活用がフィジカルAIの開発活用につながる。**
- 暗黙知を含めた現場のデータが豊富な日本こそ、バーティカルAIで独自の価値創出も可能。**
日本のバーティカルAIが国内のみならず世界の課題解決に貢献。「信頼できるAI」を創り世界をリード。

我が国の勝ち筋

主な課題 （ボトルネック）

- AI研究開発人材**、現場とAIの双方を理解して、AXを推進できる**実装人材**や**中小企業伴走支援の不足**
- 現場固有**データの活用や計算資源の不足**
- AXへの**理解不足、AIリスクへの懸念**

講じるべき施策

- 重点領域（※）を設定し、領域別戦略を策定。人材、データを含めたAI基盤への官民投資を促進。領域間の連携も促進**
- 研究開発の担い手**の育成・確保、世界から優秀な「知」が集まる機能の創設
- 中小企業**を含めた産業や行政での**AX**の促進
- 政府による率先導入と制度改革**による初期需要創出
- AIリスクへの包括的対応（AISIの機能強化や偽・誤情報対策等）
- 責任あるアジャイルガバナンス**
- 同志国・**グローバルサウス**との協働をはじめ国際連携

※ 市場性：製造、物流・交通、情報通信、金融、創薬
公共性：医療・介護・福祉、農林水産、建設、教育、行政、エネルギー
戦略性：防衛、警察、防災、消防、サイバー、海洋、宇宙、科学研究

目指すべき姿

- AIの導入で**人手不足を解消**。業務効率化、日本社会全体でのAXを実現。
- 2030年に世界のバーティカル市場は33兆円、内外で少なくとも5兆円の市場を獲得**。国内外の現場課題の解決へ。
- エコシステム全体で、AI主権を戦略的に確保。

バーティカルA I 領域別戦略 中間とりまとめ（概要）

領域設定の考え方

官民の限られたリソースを戦略的に投資し、効果を最大化することで、社会全体のA Xを実現

市場性：現場力のデータ化・集積と領域特化型基盤モデルの開発・実装により、産業構造改革と国際競争力を強化

公共性：官民連携による、A I の徹底活用を通じて、業務の効率化・負担軽減、サービスの質の向上を実現

戦略性：計算資源、データプラットフォーム、モデル、アプリ等の自律性確保により、耐遮断性と運用能力を強化

1. 市場性を有する領域別戦略：産業構造改革、競争力のあるA I システムの創出、海外展開の推進

- **製造領域：**A I ロボティクスや工場等の自動制御A I に貢献するマルチモーダル基盤モデルの開発とデータ連携手法の確立
- **造船領域：**溶接、塗装、曲げ加工等の造船の主要工程を自動化するA I 生産技術システム等の開発・導入
- **物流・交通領域：**物流領域特化モデルの開発と運行管理・遠隔監視・情報管理等に資するデータエコシステムの確立
- **情報通信領域：**通信インフラ等の現場データを連携・活用するバーティカルA I と、A I を支える通信基盤等の整備
- **金融領域：**Eージェント型A I やブロックチェーン等を活用した金融ユースケースの創出・横展開の推進
- **創薬領域：**基礎研究から臨床研究、製造工程開発等の創薬の迅速化・高度化に必要なA I 駆動型共用自律ラボ等の整備・運用

2. 公共性のある領域別戦略：A I で人手不足を解消、我が国の供給力の維持

- **医療・介護・福祉領域：**看護記録等の文書作成や福祉相談業務の支援等の現場の負担軽減のためのA I 開発普及
- **農林水産領域：**農林水産・食品産業の生産性向上のためのA I 開発・実装、データエコシステムの構築・開発基盤整備
- **建設領域：**建設機械等の協調領域のデータセット構築とフィジカルA I への活用による建設施工の自動化・省人化の推進
- **教育領域：**深い学び、多様な児童生徒への対応、教師の働き方改革を実現するためのA I の実装・データ整備
- **行政領域：**ガバメントA I を通じたA I ネイティブな業務プロセスへの変革とA I 自律性確保、行政データの機械可読性確保
- **エネルギー領域：**災害対応や設備巡視・点検、系統運用等の協調領域におけるA I 活用に係る業界内連携の推進

3. 戦略性のある領域別戦略：特定の国や企業への過度な依存を解消

- **防衛領域：**防衛力強化の観点から、データ基盤の強化やA I 実装を進め、迅速かつ的確な意思決定を実現
- **警察領域：**匿名・流動型犯罪グループによる犯罪をはじめ、様々な治安課題への対処等のためのA I 基盤の整備・運用
- **防災領域：**被災者一人一人に寄り添った支援や応急対応・復旧対策等の方針決定・遂行支援へのA I 活用の推進
- **消防領域：**災害現場等の状況を即座に分析し、迅速・的確な判断をA I が支援するための研究開発等の推進
- **サイバー領域：**「Project YATA-Shield」の推進、制御系システム等におけるサイバーセキュリティ対策強化
- **海洋領域：**海洋における異常兆候の自動検知やリスク判断を行える領域特化のA I 開発・実装
- **宇宙領域：**地上局を含む宇宙機システム全体の運用を最適化・効率化等、宇宙分野におけるA I 活用の推進
- **科学研究領域：**AI for Scienceの波及や国際・産学連携研究の推進、データ共有強化と研究インフラの一体的構築

A I 実装能力を共有する観点から、現場記録作成や意思決定支援へのA I 活用、A I ・データ基盤の開発や実装、担い手育成を省庁横断で加速

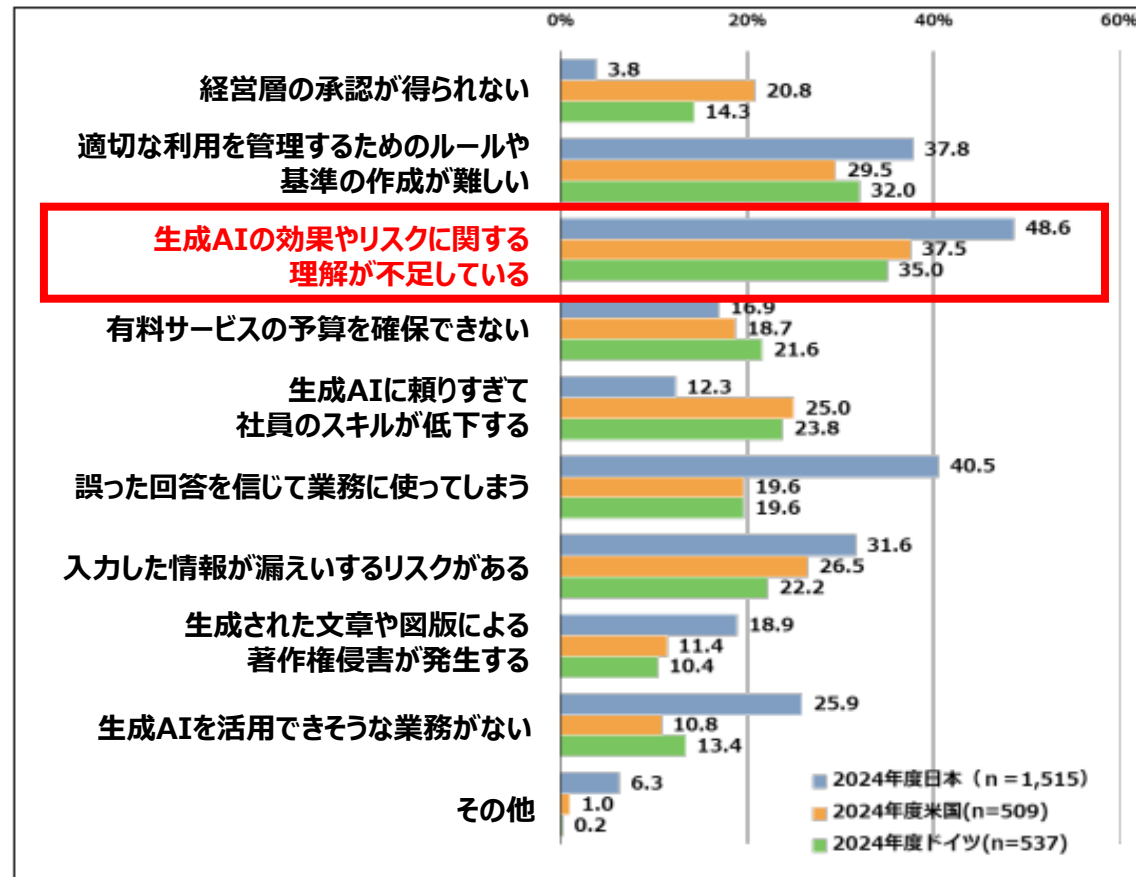
日本の生成AI利活用の課題

- 日本人のAI利活用を促進するためには、AIの必要性の理解と効果的な使い方のリテラシー教育が課題。
- 米独に比べて日本企業は、生成AIの効果が感じられず、リスクへの不安感が強い。

【個人】生成AIサービスを利用しない理由（日本）

	2024年
自分の生活や業務に必要な	40.4%
使い方がわからない	38.6%
魅力的なサービスがない	18.3%
利用環境が整っていない	8.9%
費用が高額である	3.8%
情報漏えい、セキュリティ、 安全性に不安がある	3.8%
品質に不安がある	3.7%
従来の文化や習慣を 変えられない	1.7%
その他	0.6%

【企業】生成AIを業務で活用する上での課題（日米独）



（出典）総務省（2025）「国内外における最新の情報通信技術の研究開発及びデジタル活用の動向に関する調査研究」

（出典）独立行政法人情報処理推進機構（2025）「DX動向2025」

3.2.4. 教育

①バーティカルAIを活用して目指す姿

- 深い学びの実現・多様な児童生徒への対応、教師の働き方改革を両立し、**教師が教師でなければならない業務に注力できる環境を構築**
- ガイドライン、アプリケーション、活用・実装基盤の3つの観点から取り組み、**先進事例を創出・横展開**
- 校務等の周辺業務をAIが支援しつつ、**発達段階や学習過程への影響を踏まえた安全かつ主体的な環境整備・実装により、教師の専門性を補完・拡張**

②バーティカルAIを活用していく際の課題

公共性の高さと収益性の低さ、現場の多忙さ、教育特化ユースケースの実証不足、ルール・基盤整備の遅れが複層的に絡む

市場課題

- 学校現場でのAI活用は途上であり、教育特有の要求が明確化されず、改善サイクルが回りにくい

技術課題

- 学習過程のショートカットによる学習効果の低下など教育特有のリスクへの対応が必要。また、学校現場のクラウド環境、質の高いデータの整備が急務

制度課題

- 学校現場におけるAI活用に関する留意事項や基準が十分でなく、標準・共通仕様の策定が進みにくい

人材課題

- 教育委員会や教職員のAIリテラシーに差があり、多忙な現場では試行・改善の余力が乏しく、事例の横展開が進みにくい

③課題解決の手段・アプローチ

ガイドライン、アプリケーション、活用・実装基盤の一体的な整備を通じ、GIGAスクール構想・次世代校務DXを推進

- ガイドラインを速やかに改定し、研修等を通じて周知するとともに、AISIに設置される教育WGで評価基準等を検討
- 校務領域では、学校運営や外部対応等をAIが支援し、BPRを通じて事務負担軽減と働き方改革を進める
- 深い学びの実現、個別最適・協働的な学びの一体的な充実、多様な児童生徒への対応に資する教育分野特化AIの実証・実装を進める
- 児童生徒の利活用では、発達段階や学習過程への影響、ハルシネーション等のリスクを踏まえ、実装を検討する必要。
- AI実装に不可欠なセキュリティ等のインフラと、公共性の高いデータを整備

④2030年に向けて取り組む主な施策

ガイドラインを今年度中に改定するとともに、いわゆるガードレール付きAIなどを含む教育分野特化AIの実証、質の高い教育データ、クラウド基盤、教職員研修、英語教育へのAI活用に官民投資を集中し、学校現場での安全かつ主体的なAIの利活用を加速する

①ガイドラインの改定

- 学校現場における利活用のガイドラインについて、学習指導要領の改訂を待たずに令和8年度中に改定するとともに、その後も技術の進展等に応じて対応。

②教職員向けの研修実施

- ガイドラインを周知しつつ、AIの仕組みや特徴への理解を深め、学校現場での適切な利活用を推進するための研修を全国的に実施する。まずは、2026年度中に3,000人規模の研修を実施

③教育分野特化の実証研究事業

- 校務におけるAI活用を通じた働き方改革、AI活用を通じた多様な子供への対応、深い学びの実現に向けたAI技術の活用、新たな技術の検証を一体で実証

④AI活用による英語教育の強化

- 英語教育においてAIを効果的に活用するため学習指導要領を改訂するとともに、効果的なAI活用に関する好事例を創出

⑤教育分野における質の高いデータの整備

- 学習指導要領、参考資料、教科書、MEXCBT搭載問題等の質の高い構造化データの整備に向けた事業を実施

⑥主体的かつ安全にAIを活用できるクラウドベースのインフラ整備

- AIを安全に活用できる学校現場における必要なセキュリティ確保等のための支援

日本AISIと英国AISIの違い

	日本 AI Safety Institute 	英国 AI Security Institute 
役割	国内エコシステム全体を動かす司令塔・共通基盤	政府内部に高度なAIモデルの評価能力を保有
設立	2024年2月	2023年11月
所管	経済産業省・デジタル庁 IPA（情報処理推進機構）内	Cabinet Office内組織 （26年7月から科学・イノベーション・技術省から移管）
所長	村上明子	ヘンリー・ドウ・ゾエット(Henry de Zoete)
職員数・予算	61人*（併任含む） 令和7年補正：88億円（IPA運営費交付金：25億円） * 令和8年9月時点（4月時点で31名）。	約200人以上*（うち専門家は90人）（目標300人） 予算：約130億円、そのほか、官民共同基金約60億円や現物提供などの民間からの支援もあり *2025年9月時点、大学教授、元Google、元Open AI等トップ人材を直接採用。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
主な機能	AIの社会実装全体を対象 注：モデル評価だけでなく社会影響、ガバナンス、AIシステム、コンテンツ、データ標準など広範な範囲に対応 - 政府AIサフェティ評価観点ガイド、レッドチーミング手法ガイド、AIサフェティ評価ツール開発・提供 - 政府のAIガイドライン作製への協力 - 産業別AIガイドライン作製（ロボ、ヘルスケア他） - AI安全性に係る国際標準・認証との接続支援 - 国際連携	高度なAIがもたらすリスクについて政府に科学的理解を提供 - 主要なAIモデルを公開前・公開後に評価 - フロンティアAI企業と連携して安全性を改善 - 評価基盤開発・公開（評価ツールを提供するサイト運営） - 内部での技術研究と研究助成 - 英国政府及び同盟国政府へのリスク情報の提供 - 政策への脅威情報の提供
AIモデル評価に係る立ち位置	事業者や評価機関が評価できるよう、評価方法や物差し・ツールを提供	フロンティア企業から公開モデルへのアクセスを受け、自身でAIモデルをテストし、英国政府や同盟国に共有
権限	規制・認証・行政処分を直接行う機関ではない	規制当局でなく研究・評価機関（将来的に法定組織化も構想）
組織モデル	産総研やNICT等の研究機関とのネットワーク・官民連携組織	内製研究所型・技術集約型

→日本AISIは英国AISIをコピーするのでなく、信頼できるAIの実装の本格化に向けて、現在の広範な社会実装・標準化機能を維持。そのうえでUK型の技術評価部門を組み込むことが必要（①技術評価人材の増強、②フロンティアAIモデルへの事前アクセス権確保、③政府調達・重要インフラ向けの独立評価能力、④サイバー、CBRN、フィジカルAI安全保障評価の内製化、⑤AISIの技術的知見が政府に直接届く仕組みの強化）

AISI事業実証WGへの期待

- 高性能AIモデルを評価できる能力としては国として必要。
- 政府が目指す「信頼できるAI」の社会実装や、バーティカルAIやフィジカルAIの展開において必要なことは抽象的な議論ではなく、**「どう評価し、どう安全性を担保するか」という技術的・実務的な物差しを作ること**
- それを出来るのはこの場にいる皆さん
- **AISI自体の機能強化・体制強化が進む中で、日本ならではの強みを構築していくために、ここでの議論を基に日本発の評価方法やベストプラクティスを世界に向けて発信していきたい**

■ 経済財政運営と改革の基本方針2026（令和8年7月21日）（抄）

「信頼できるA I」で社会全体を駆動するA Xに国を挙げて取り組む。このため、デジタル行財政改革会議を改組し、人工知能戦略本部、規制改革推進会議等と一丸となって、A Iの開発・利活用に係る法制度やガイドライン等について能動的かつ抜本的な見直しを行うなど、省庁横断でA X／D Xを実行する。

「世界で最もAIを開発・活用しやすい国」をスローガンに、官民が一体となって、日本のAIの未来を創り上げていきましょう