資料１-１

人工知能戦略推進会議幹事会

AISI機能強化の方向性（案）

令和8年４月７日

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。AISIの安全性の評価能力を確立**しながら、**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。

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。

自民党デジタル社会推進本部　AI・web3小委員会提言(R7.12.19)）
～ＡＩセーフティ・インスティテュート（ＡＩＳＩ）の機能強化に係る緊急提言～

我が国のＡＩに関するイノベーションの促進とリスク管理を両立させるためには、**「信頼できるＡＩ」の利活用及び開発の中核となるＡＩセーフティ・インスティテュート（ＡＩＳＩ）の抜本的な機能強化**を行わなければならない。**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の構築**を行う必要がある。

このＡＩＳＩの機能拡充及び機能強化においては、**政府を挙げた取組み**が必要であり、特に以下の二つの目標を早急に達成しなければならない。

まず、世界のＡＩ開発事業者から、フロンティアモデルの発表、提供に先立ち、事前評価の実施を委託される機関となる。当面は他の独立行政法人や民間機関等との連携の下、将来的には自ら、技術評価能力の強化とそのための研究開発基盤を構築する。**世界の主要開発事業者との協力協定を積極的に締結**する。

また、顕在化する「ＡＩによるサイバー攻撃とＡＩによる防御」に対応できるよう、諸外国のＡＩＳＩや内外の関係機関と連携し**サイバーセキュリティの評価機能を強化**する。サイバーセキュリティに関する専門人材をはじめ人的基盤を強化する。

ＡＩＳＩを軸とした日本として安全性やセキュリティ確保に係る国際ネットワークをグローバルサウスを含めて構築し、AIサミットの日本での早期に開催も行うことで、日本の「信頼できるＡＩ」を世界に拡げていく。

そこで、ＡＩ・ｗｅｂ３小委員会・デジタル社会推進本部として、ＡＩＳＩの機能強化について、下記のとおり緊急提言する。

自民党デジタル社会推進本部　AI・web3小委員会提言(R7.12.19)
～ＡＩセーフティ・インスティテュート（ＡＩＳＩ）の機能強化に係る緊急提言～

1. 政府は、**英国のＡＩＳＩをベンチマークに、質・量ともにＡＩＳＩの人員・体制強化**を図ること。まずは令和7年度補正予算を的確に執行し、**早急に現行の30名から陣容を拡充し、60名体制を目指すこと。**

2. ＡＩはあらゆる行政分野に関係しており、**各省庁でＡＩ安全性やＡＩセキュリティの専門家の育成が必要不可欠**である。そこで、**全省庁がＡＩＳＩに出向者を出す**こと。**特にデジタル政策に関係する省庁については複数名**出すこと。

3. ＡＩＳＩが自らの権能で国内外の有能なＡＩ関連の専門家を柔軟に雇用できるよう、国家公務員より高額かつ柔軟な年俸支出も可能となるようにすること。

4. ＡＩＳＩの拡充する業務の適切な執行と常勤常駐含めた体制整備及びそのための財政基盤を確保するため、**ＡＩ政策の司令塔である内閣府がＡＩＳＩ業務の共管省庁となる**こと、また**内閣府及び経済産業省から運営費交付金を安定的に支出**すること。

5. ＡＩサミットの日本での早期開催を検討すること。

人工知能戦略本部　総理ご指示（2025.12.19）

ＡＩは、産業競争力や安全保障に直結。信頼できるＡＩによる日本再起を実現するため、以下を指示（以下、抜粋）

- 第一に、『ガバメントＡＩ源内』の徹底活用です。2026年５月から１０万人以上の政府の職員が活用できるようにします。ＡＩ源内の活用により、創造的に業務を行い、国民の皆様に信頼できるＡＩの意義を示してください。

- **第二に、ＡＩセーフティ・インスティテュートの抜本的強化**です。ＡＩの安全性に対する不安が高まる中、**英国並みの２００人体制を目指し**て、小野田大臣と赤澤経済産業大臣は、**全省庁、産学から人材を集結させ、ＡＩセキュリティに万全を期してください**。

- 第三に、ＡＩロボットを始めとしたフィジカルＡＩに不可欠な信頼できる国産の汎用基盤モデルの開発です。赤澤経済産業大臣は、質の高い産業データを日本の競争力の中核に位置づけ、意欲ある企業としっかりと連携し、開発を進めてください。

- 第四に、信頼できるＡＩによる社会課題を解決できるサービスの開発・導入です。今般の経済対策で、４０００億円以上のＡＩ関連施策を措置したところです。これらを活用して、地域や中小企業の成長戦略を実現するとともに、世界各国にサービスを展開してください。

- 第五に、信頼できるＡＩを世界とともに創りあげるため、『ＡＩサミット』を可能な限り早期に日本で開催すべく、関係省庁を挙げて、取組を進めてください。

- 第六に、信頼できるＡＩを創る官民投資を日本成長戦略における危機管理投資として、力強く推進してください。政府としては、投資の予見性を高めるため、当面、１兆円超をＡＩ関連施策の推進に投資してまいります。また、大胆な投資促進税制を創設し、研究開発税制を深堀りします。これらの政府のコミットを、それぞれが所管する企業の皆様と共有し、政府の取組に呼応していただきＡＩ投資を強力に推進してください。

- 結びに、ＡＩをめぐる動向の変化は非常に速いです。小野田大臣は、今回の計画に基づく、官民の取組を直ちに実施するとともに、来年の夏を目指して、投資目標、制度改革、人づくり、データ戦略などを含む官民投資ロードマップを盛り込む形で、『ＡＩ基本計画』を更に充実させてください。以上です。

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)**

AISIへの職員等の出向

令和８年２月中旬から、必要省庁に人員派遣を依頼中

注：党の提言等を受けて、一部省庁等から相談が来ており、並行して事前相談中。

■AISIへの職員等の出向に係る方針

ＡＩＳＩでは、高市総理からの指示及び党からの提言を踏まえて、早急に現行の30名体制から60名体制への移行を進めるべく、研究者を中心にＡＩ関連の技術者の雇用を進めている。
政府を挙げてこの取組を後押しするため、AISIからの要望を踏まえつつ出向者の検討、調整を行う。

1. 出向者は、令和7年度補正予算の執行、及び高市総理指示及び自民党提言への対応を行うため、以下の業務を行うことを想定する。**業務を行える者であればクラスを問わないが担務としては補佐又は係長級を想定（合計13名（調整中））**。
   - ① ＡＩＳＩの組織運営業務（企画・総括、会計・契約、人事、組織の規程改訂、国や民間企業との渉外対応等）3名
   - ② 海外ＡＩＳＩ機関等との連携・協力・調整業務　１名
   - ③ ヘルスケア、ロボティクス、金融、教育等の国の重要な分野におけるＡＩの安全性評価ガイドライン策定、データ品質管理、AI認証制度の検討に係る業務　５名
   - ④ ＡＩセキュリティや安全保障対応に係る業務　３名
   - ⑤ 不適切なAI開発・利用実態把握等を含めたAI利用実態調査の業務　１名
2. 指示や提言への早期対応のため、各省庁は、**早ければ令和８年4月から、遅くとも7月頃までに、出向者を出すことを調整する**。出向期間は１～2年程度を想定する。

AISIの機能強化の方向性（案）について

１．これまでの機能強化の方向性に係る検討状況

- 3月３日　基本計画推進WG①（3月25日に個別に欠席委員に説明）
- ３月5日　産総研、理研へのヒアリング
- ３月9日　NICTへのヒアリング
- 3月11日　NIIへのヒアリング
- 3月19日、27日　AISI幹事会（関係省庁）
- 3月30日　基本計画推進WG ②

※その他、サイバーセキュリティに係るAISIの取組について、NCO主催の「サイバーセキュリティ推進専門会議」でも相談

２．主なコメント

- AISIのリソースやタイムラインを踏まえて、**優先順位を付けて今後の機能強化の具体化**を図るべき
- **各機関の多様な研究を維持しつつ、AISIを中心にいかに全体の機能強化**
- AISI等と連携した形でのAIセキュリティに関する官民での連携強化は重要。**海外で発生したAIインシデントや既存のレポートラインに乗らない情報**は、AISI国際ネットワークも活用しつつ、**情報収集・共有いただきたい**
- **国際協調に係る対外的なポジショニングを検討する場**がないのは気になっている。
- **各パートナーシップ機関が率直に議論できる場**を継続的に設けてほしい

AISIの機能強化の方向性（案）について

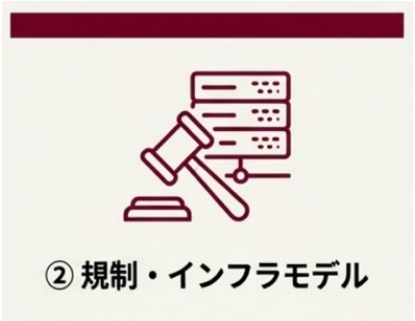
3タイプある世界のAI安全保障戦略のいいとこどりを

関連のパートナーシップ機関や民間企業、政府機関と連携し、
「官民共創」の安全ベンチマークで「自己評価をできる能力」を持ち、
必要に応じて規制するための**ソフト・ハードロー作成の技術支援**を行う

日本モデル
官民エコシステムで作成したベンチマークを用いて、政府が評価能力を持って政策を実施

英国・米国モデル
政府が直接評価能力を内製化し技術的優位を確保

Measurement Science and avoiding surprise



EU（フランス）モデル
AI法と市場のルール構成

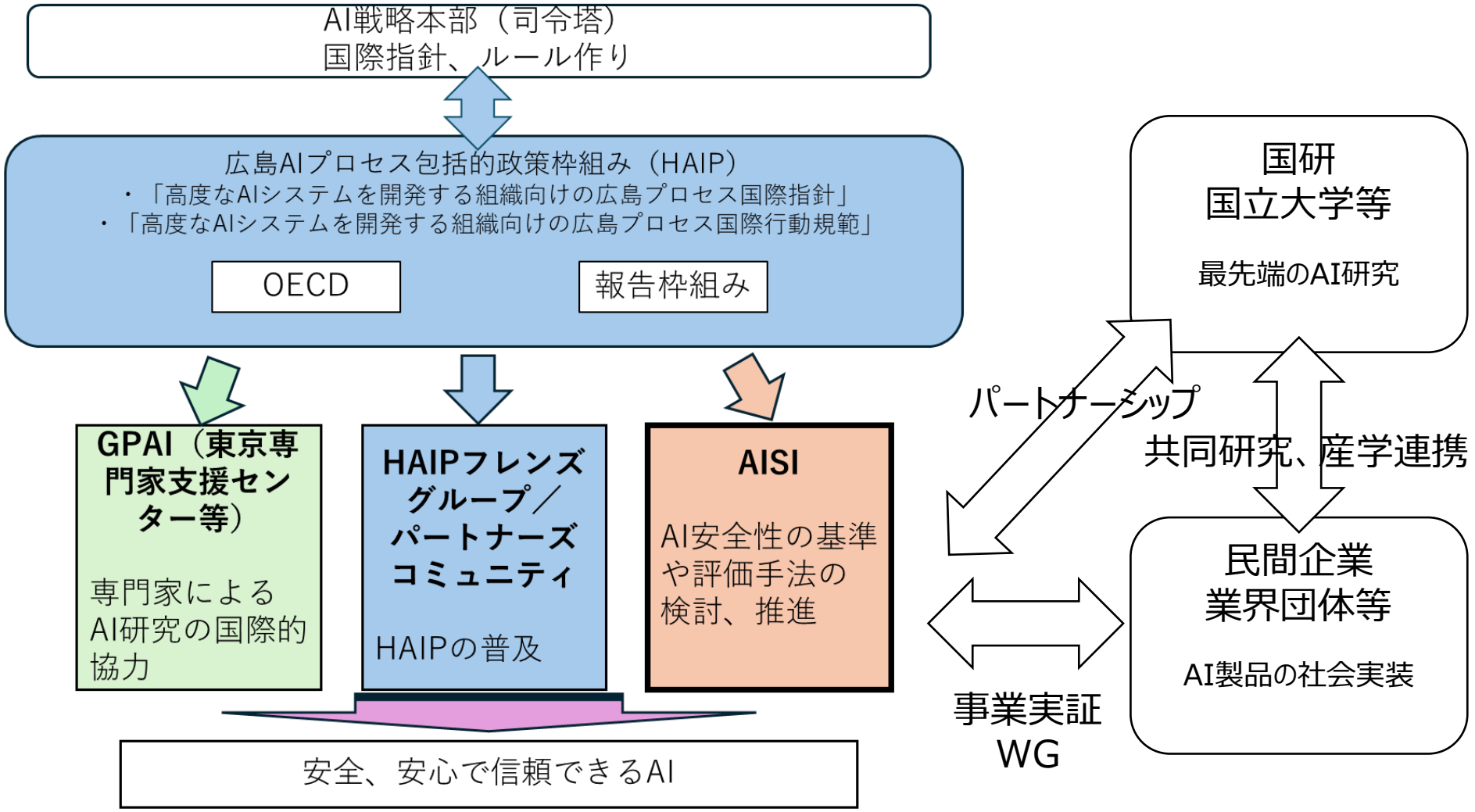
Regulatory Enforcement and Update based on Trust

シンガポール・カナダ・韓国モデル
オープンソースと研究コミュニティの力で社会実装を主導

Implementable Guidance and Leveraging the Research Ecosystem

AISIの機能強化の方向性（案）について

（参考）パートナーシップ機関やGPAIとの関係、役割分担



AISIの機能強化の方向性（案）について

● ミッションの再整理

「AISIのパーパス：「信頼できるAI」の提供に必要な情報や技術を有し、日本を世界で最もAIを開発・活用しやすい国にする」

パーパス実現のために、AI基本計画等を踏まえて、**ミッション（AISIの役割・責任）を整理**

①AI安全性に関連する情報のハブとして、**信頼できる優れたAI活用製品やサービスの普及**（もって、国民生活の向上・社会課題の解決を促進）

②AI技術やAI安全性等に係る国際標準化等により、**我が国のAI関連分野のイノベーションの加速・競争力強化**

③情報収集・分析・共有・対策の早期実施により、**AIの開発・普及に伴う国民の生命・財産に危害が及ぶような事象など、国家の安全を脅かす事象の抑止**

● ミッション達成に必要なAISIが行うべき４つの活動

（１）AIの物差しを作る
（２）自ら評価する能力を持つ
（３）AI関連情報の収集・分析・提供
（４）国際協調の主導

AISIの機能強化の方向性（案）について

● これまでのAISIの活動と、今後AISIが行うべき活動

（１）AIの物差しを作る

①AIの安全性の確保を促進するガイドラインの作成・提供

（これまで）
業界横断的かつ優先度の高いガイドラインの作成を先行実施。

・AIセーフティ評価観点ガイド
・レッドチーミング手法ガイド

（今後）
- マルチモーダルAI、AIエージェント、フィジカルAIなど、**進化する最新AIに対応したガイドラインの作成、既存ガイドラインの改訂**
- **個別の産業分野特有の課題に特化したガイドラインの作成**（ヘルスケア、ロボティクスなど）

注：「Constitutional AI」のようなAIが判読するようなものも検討

②AIの安全性を評価する手法の開発・提供

（これまで）
・AIの安全性を評価するための評価環境の開発に着手。
・ベンチマーク設計方針の検討、サンプルデータの整備を行うベンチマーク作成コンソーシアムが発足。

（今後）
- 進化する最新AIを評価するための**ベンチマーク構築・改良、データセットの充実化**
- **産業分野毎のベンチマーク開発**（ヘルスケア、ロボティクス、金融、教育、・・・）
- **最新のAI評価技術の試験的利用**（ NICT等のパートナーシップ機関と連携）

- ①、②共通：ガイドラインや評価手法について**民間企業や政府機関からのフィードバックを基に改善**

AISIの機能強化の方向性（案）について

- **これまでのAISIの活動と、今後AISIが行うべき活動**

（１）AIの物差しを作る

③各AI製品・サービスの安全性が適正に評価され、その情報が流通する仕組みの構築（第三者認証制度など）への貢献

（これまで）
・ISO/IEC 42000シリーズ策定に際して技術面での助言
・ISO/IEC 42001のJIS化に向けた貢献

（今後）
- AISTとともに、AIに係る我が国発の**国際標準の策定を推進**
- 国際標準の実用化に向けたコンフォーマンステスト**（製品やサービスの適合性試験）の検討**
- ベンチマーク作成コンソーシアムで得られた知見を基に、**第３者AI認証制度への技術提供**

AISIの機能強化の方向性（案）について

● これまでのAISIの活動と、今後AISIが行うべき活動

（２）自ら評価する能力を持つ

①　AIによるサイバー攻撃・防御への対応（技術動向の調査・共有等）

（これまで）
・AI脆弱性情報、インシデント情報の収集
・セキュリティレポートの検討
・サイバーセキュリティに関する勉強会の開催

（今後）

- **<u>AIインシデントに係る情報※の収集及びその評価</u>**
  政府関係機関（NCO、NSS、防衛省、経産省等）と共有、対応策について官民が連携して取り組む体制の検討

※AISI国際ネットワーク等の枠組みも活用しつつ、海外で発生したAIインシデントや既存のレポートラインに乗らない情報等を想定（サイバー対処能力強化法に基づく官民協議会も活用）

②　新たなAIについて、影響の大きな悪用ができるかどうかを評価

（今後）

- AIの第三者評価が可能となる**<u>評価環境構築の本格化</u>**（IPAやNICT等のパートナーシップ機関との連携を図る）
- 構築したAI評価環境による**<u>フロンティアAIの悪用評価</u>**（MoC締結企業の新製品評価、LLMの評価からAIエージェント、フィジカルAIの評価へと拡張）

AISIの機能強化の方向性（案）について

● これまでのAISIの活動と、今後AISIが行うべき活動

（３）AI関連情報を収集・分析し、提供する

① AIの安全性向上に資する関係者間の情報共有の促進と政府の政策への技術的観点からの助言

（これまで）
・AIセーフティに関する事業実証を行うWGを設置（ヘルスケア、ロボティクス）
・ベンチマーク作成コンソーシアム内での情報共有

（今後）
- 事業実証WGの実証結果を踏まえ、官民が連携した**業界毎のAIガイドライン、制度設計への技術的助言**（NICT等のパートナーシップ機関と連携を図る）

②誤動作等が発生した場合の影響が大きい分野のAI事例の情報収集、分析

（今後）
- **影響が大きな分野**（例：電力網などインフラ分野、モビリティ）**へのAI導入事例や、AIインシデント事例につき幅広に情報収集、分析**
- 業界別ガイドライン、産業分野毎ベンチマークへの反映

AISIの機能強化の方向性（案）について

- **これまでのAISIの活動と、今後AISIが行うべき活動**

（３）AI関連情報を収集し、分析し、提供する

③影響の大きいAIの悪用や誤動作の可能性があるAIの開発・流通を防止する政策的対応への助言

（今後）

- **AI法16条調査等への技術的支援**（ディープフェイク、知的財産権侵害、誤情報やバイアスによる教育・人事・採用等における差別的な扱い等、国民の権利利益の侵害について影響の大きいAIリスクとその対抗策を共有）
- AI開発事業者との連携により、製品中のAI評価を行い、**リスクの高いAI製品の流通を防止するための技術的情報を関係者に提供**

④不適正なAIの開発・利用の国際的な事案の把握、政府への情報提供

（今後）

- **AI脆弱性情報・インシデント情報の収集・分析**、AISI国際ネットワーク（英、米、カナダなど）相互の情報共有、OECDインシデントレポートへの貢献、
  国際的に重大なインシデント発生時に対応可能な官民連携体制構築
  （官：NCO、NSSなど　民：AI開発企業（OpenAI、Anthropic等）、AI利用企業）

AISIの機能強化の方向性（案）について

- **これまでのAISIの活動と、今後AISIが行うべき活動**

(4)国際協調を主導

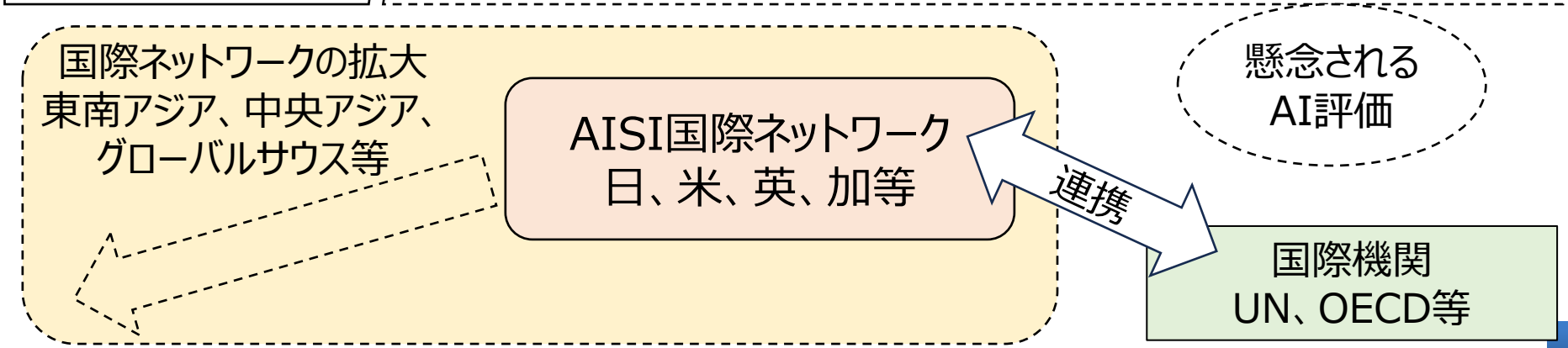
(1)～(3)の活動に資する海外機関との連携・協力

（これまで）
・AISI国際ネットワークの構築
・HAIPを端緒とするAIサミットへの参加
・AISIネットワークイベントの開催
・JICAへの講師派遣

（今後）

- **<u>AISI国際ネットワーク間の連携強化、ベストプラクティスの共有</u>**
- **<u>国際ネットワークの拡大</u>**（新興国のキャパシティビルディングへの技術的協力）、国際機関との連携（OECDツールカタログ、ソリューションカタログなど協働プロジェクトに貢献）
- **<u>懸念されるAIの評価により国際的な対応をリード</u>**

注：関係省庁に伺いながら国際協調に係る方針を検討



AISIの機能強化の方向性（案）について

（その他の機能強化に向けた取組）

- 研究者が協力しやすく、キャリアップにつながるAISIの組織体制を構築。パートナーシップ機関と協力したコミュニティの造成を実施。
- **サイバーセキュリティについては、NCOやデジタル庁等と相談しながら、安全保障に係る取組については、NCOや防衛省と警察庁等と相談**しながら、具体の機能強化を進める。
- これらの取組の実効性を確保するため、セキュリティ・クリアランス制度も活用し適切な情報管理がなされるよう取り組む。
- **パートナーシップ機関との間で活動の意見交換をする場を設定**。また、パートナーシップ機関等の国内施策をマッピングした上で、AISIにおいて不足部分を検討（要すれば、パートナーシップ機関との協定の記載内容も見直し）。
- 別途、今春中を目途に、党からの提言を踏まえて、
  内閣府がAISI業務の共管省庁となる（法改正が必要）、
  内閣府及び経済産業省等からAISIに安定的に財源を支出する、
  ための検討が必要。

（参考）AISIの機能強化の方向性（案）

● 今後AISIが行うべき活動（全体像）

	これまで	今後（R8年度から段階的に強化）
AIの物差しを作る	• AIセーフティ評価観点ガイド • レッドチーミング手法ガイド • データ品質マネジメントガイドブック • CAIOガイドブック	• ガイド等の更新 • 最新AI対応ガイドライン・基準 • ベンチマーク • 事業実証WG・コンソーシアムの業種拡大 • 第三者認証制度への提言 • そのために必要な個別研究・技術開発
自ら評価する能力を持つ		• フロンティアAIの悪用評価 • 影響が大きな分野でのAIエージェントやフィジカルAIなどの評価 • 国際的に懸念されるAIの評価
AI関連情報を収集、分析、提供する	• AIインシデントレスポンス・アプローチブック • セキュリティ攻撃に関する詳細レポートの公開	• AIインシデント情報の共有 • 重大AIインシデントへの官民連携体制構築 • AI法16条調査等の支援
国際協調する	• AISI国際ネットワーク参画 • JICA国際研修への派遣	• AISI国際ネットワーク連携強化（ベストプラクティスの共有） • 諸外国およびAI企業との協働プロ

（参考）AISIの機能強化の方向性（案）

● AISIの機能強化に向けたロードマップ

	～１年（R８年度）	～３年（R１０年度）
体制	60人体制拡充	100人体制拡充
AIの物差しを作る	WG検討対象業種の拡大 ベンチマーク等検討 → ベンチマーク開発・ガイドライン改訂（最新AI（AIエージェント）、産業分野別） 認証スキーム検討 → 認証制度化 標準化検討議論（国内） → 国際標準化提案	
自ら評価する能力を持つ	AI評価環境構築 → AI評価実施 関連調査 → 化学・生物兵器製造、ディープフェイク、著作権侵害等の評価	
AI関連情報を収集、分析、提供する	セキュリティ情報共有体制検討 → セキュリティ情報収集・共有・対処	
国際協調する	AI開発企業との連携 → その他国内外AI関係企業との連携拡大 関係国と協働プロジェクト検討 → 協働プロジェクト実施	

AI agentの製品実装

AGI、汎用AIの実装

参考資料

AIガバナンスを巡る国際動向

G7及びOECD等

G7広島サミットにて、日本が広島AIプロセス提唱

安全、安心で信頼できる高度なAIシステムの普及を目的とした初の国際的政策枠組みとして**「広島AIプロセス包括的政策枠組み」**がG7首脳で承認

広島AIプロセスのアウトリーチとして、**広島AIプロセス・フレンズ・グループ創設⇒グローバルサウス等への働きかけ開始**

OECD閣僚理事会 生成AIに関するサイドイベントの実施、**OECD AI原則の改定**

広島AIプロセス・フレンズ・グループの**初の対面会合が開催**

広島AIプロセスに即した指針の整備を規定した**AI法が公布**

(2023年) (2024年) (2025年) (2026年)

5月 11月 12月 2月 5月 2月 6月 12月 2月

AIサミット
AI安全性サミット
（英国で開催）

AISI
「AI安全性」を具現化するための議論が進み、英・米はAI Safety Instituteを設置

AIサミット
AIソウル・サミット
（英・韓共催）

AISI
日本のAI Safety Instituteを設置

AIサミット
AIアクションサミット
（仏で開催）

AISI
英国AISIの名称がAI Security Instituteに変更

AISI
米国AISIの名称がthe Center for AI Standards and Innovation (CAISI)に変更

AIサミット
AIインパクトサミット
（印で開催予定）

パックス・シリカ・サミット
AI関連サプライチェーンの強化等をテーマに米国主催で開催された新たなサミット

AIサミット及びAISI等

AIを巡る国際動向と日本の役割

日本のAI法及びAIガバナンスへの取組への期待は高い

非規制
輸出促進

自主的対応
の促進
（ソフトロー）

リスク対応
の法規制
（ハードロー）

米国

- 2025年7月、**「AI競争に勝つ：アメリカのAI行動計画」**を発表。①AIイノベーションの加速、②AIインフラの整備、③国際的な外交・安全保障での主導の3本柱で構成する包括的国家戦略。イノベーション促進と米国発モデルの国際標準化**（AIスタック輸出、オープン化）を重視**。
- また、同年7月の新たな大統領令において、連邦政府が調達するAIモデル、特にLLMに対して、真実性とイデオロギー中立性を求めることに**（過剰にリベラルなAI（Woke AI）の排除）**
- 同年11月、世界で最も強力な科学プラットフォームを構築するための**国家的なイニシアチブ「Genesis Mission」**を発表。**10年以内に米国の研究とイノベーションの生産性と影響力を倍増**させる

日本

- **2023年G7議長国下で「広島AIプロセス」を立上げ**、安全・安心で信頼できるAIを実現するためのルール作りを推進。「国際指針」と「国際行動規範」を含む「包括的政策枠組」を策定。フレンズグループを通じて理念の共有を進め、参加国・地域が56に拡大。
- **2025年5月にAI法が成立し、イノベーション促進とリスク対応のバランスを追求**。
- 同年12月でAI基本計画及び適正性確保のための指針を策定。個人の尊厳が尊重される人間中心のAI社会を堅持しつつ、**「信頼できるAI」を追求し、「世界で最もAIを開発・活用しやすい国」の実現**を目指す。

EU

- **基本的人権の保護を重視。欧州AI法に基づく規制**（リスクの程度によりAIを4つに分類し、それに応じて禁止事項や義務を規定）と並行して、イノベーションも推進。
- 2025年4月、欧州委員会は、**欧州がAIの世界的リーダーとなることを目的として「AI大陸行動計画」**を発表。**①AIデータ・インフラ構築、②高品質データへのアクセス拡大、③イノベーション促進と戦略分野導入加速、④AI人材強化、⑤規制遵守の促進・簡素化**の5つの柱から構成。
- 同年11月に、欧州委員会は、**「高リスク」のAIに関する規制の導入時期を16か月遅らせることを発表**

国家管理

中国

- 2024年9月、国連の「グローバル・デジタル・コンパクト（GDC）」のビジョン実現のための「AIに係る能力構築に関する国際協力に向けたフレンズグループ」を立上げ。
- **2025年7月**、世界AI大会を開催。**「AIグローバルガバナンス行動計画」**を発表。**「世界AI協力組織」（本部：上海）の設立を提唱**。
- 「行動計画」では、国連の下でのガバナンス・メカニズムの設立等を含む13項目が盛り込まれ、**国際協力の重視を強調**。
- **2025年8月、「『AI＋』行動のさらなる実施に関する意見」**を発表。2035年までの三段階目標を掲げ、AIを社会・経済全域に深く融合し新質生産力と知能社会を育成する行動提言。

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）によりシフト。
防衛科学技術研究所等とも連携。

- 2025年6月3日、**米国**AISIはCAISI（Center for AI Standards and Innovation） に改名。
これまでの**AI安全性に係る取組から、より競争力強化（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にフォーカス。**具体的には、
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）。
米国は、日本や韓国との間でMOUを結び、業界標準の開発等を推進。

- 2025年7月26日、**中国**は、上海での世界人工知能大会で、AIへのユニバーサルアクセスが多くの国で確保されることの重要性を説く中で、グローバル・サウスへの支援を表明。
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。
(参考）同国サイバーセキュリティ法を改正し、AI等の新技術のサイバーセキュリティへの活用（AIによる攻撃を想定しAIによる防御を実施）を追加予定。

- 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**サイバーセ攻撃等に対する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」)

- **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。
- **AIは防御にも不可欠**。検知技術の向上や安全対策の強化がますます重要に。

【AIを活用したサイバー攻撃の進化状況】

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**。

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

日本AISIの現状に対する内閣府の評価

- **主要国AISIのように、技術評価のR＆Dリソース・体制が十分になく、自らAI評価ツールを作ったり、評価することができていない。**
  注：**他国AISIでは専用テストベッド・計算資源を準備**、フロンティアAIモデルの事前評価を実施。

- 日本の国研には**AI安全性の知見・ノウハウが存在するも、AISIがこれら国研の司令塔にはなれていない。**現在は、パートナー協定を結んでいる国研（NICT、産総研、NII、理研、IPA等）の成果を活用するのみ。最近、民間企業への外注を開始。

  例１：NIIが日本語LLMの出力の安全性・適正性向上のためのデータセットの作成等を実施、それをAISIの名で各国に提供
  例２：NEC等に外注して、AIが有害な情報を出力しないか等を自動で評価できるツールを開発、AISIの名で公開

- **国家安全保障やサイバーセキュリティの観点から、AIの意図的な誤動作や思想誘導等の評価や、バックドアがないか等の分析は行っていない。**

- **この背景としては、人員・予算が潤沢でないため。**
  結果的に、企業向けのAI安全性に係るガイドのみを作成・提供。

  例：AI安全性に関する評価観点ガイドやレッドチーミング（攻撃者側の目線でAI脆弱性を発見する手法）ガイド等