資料３－２

AIセーフティ・インスティテュート(AISI)
概要と2024-25成果

2026-1-22
AISI事務局

AISI Japan AI Safety Institute

AISIの役割とスコープ

AISI Japan AI Safety Institute

AIの安全安心な活用が促進されるよう
官民の取組を支援することがAISIの役割

役割

- 主に3つの役割を担う。

政府への支援
- AIセーフティに関する調査、評価手法の検討や基準の作成等

日本におけるAIセーフティのハブ
- 産学における関連取組の最新情報の集約
- 関係企業・団体間の連携促進
- 他国のAIセーフティ関係機関との連携

関連の研究機関との連携実施
- 国研等の関係研究機関との連携
- パートナーシップ事業の推進

AIの開発や利用をする者が**AIのリスクを正しく認識**できる仕組みの構築 ＋ ガバナンス確保などの必要となる対策を**ライフサイクル全体で実行**できる仕組みの構築 ⇔ **国内・国際的な関係機関**

▼

イノベーションの促進と**ライフサイクルにわたるリスクの緩和**を**両立**する枠組みを実現

スコープ

- AIによる以下の事象や検討事項の中で、諸外国や国内の動向も見ながら柔軟にスコープを設定し取組を進めていく。

社会への影響	ガバナンス	AIシステム	コンテンツ	データ

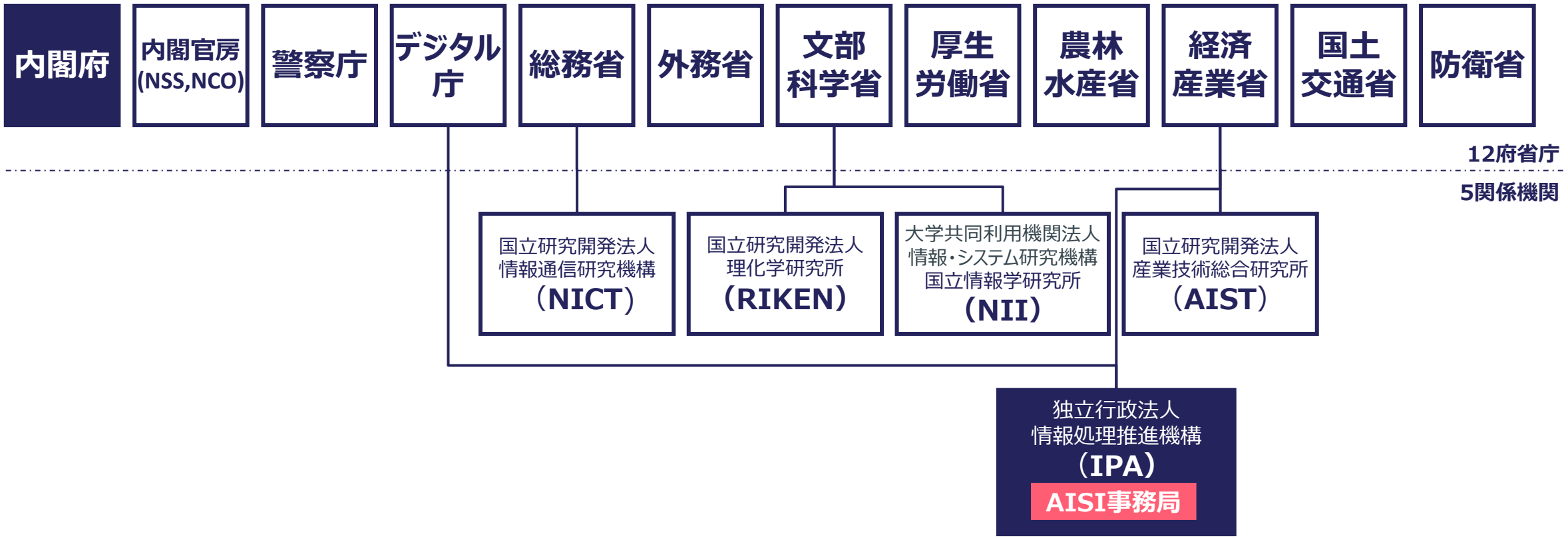
AISIの関係府省庁・機関

AISI Japan AI Safety Institute

AISIは、12府省庁・5関係機関が横断的に参画する**政府関係機関**

事務局は経済産業省とデジタル庁を所管官庁としている**IPA内に設置**

＊2025年4月時点



AIセーフティ実現に向けた業務

AISIは、**安全性評価とその実施手法**に関する検討や、
国際連携に関する業務などを遂行

1. 安全性評価に係る調査、基準等の検討

- 安全性に係る標準、チェックツール、偽情報対策技術、AIとサイバーセキュリティに関する調査
- 安全性に係る基準、ガイダンス等の検討
- 上記に関するAIのテスト環境の検討

2. 安全性評価の実施手法に関する検討

3. 他国の関係機関との国際連携に関する業務

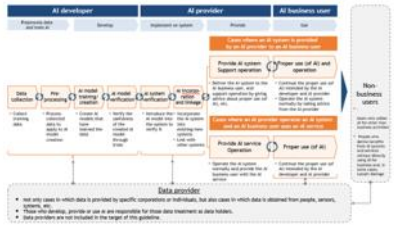
成果物の概要

AISI Japan AI Safety Institute

AIの安全安心な活用促進に向け、様々な成果物を公表

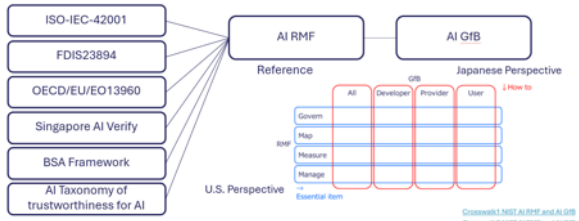
AI事業者ガイドライン

総務省と経産省が策定・更新



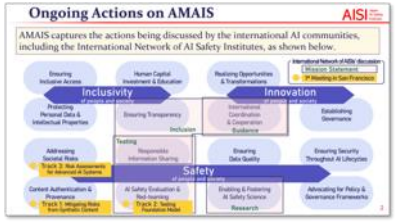
クロスウォーク

国際的な相互運用性の確認



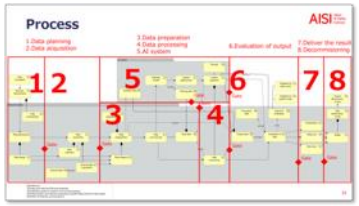
活動マップ

AI安全性に関する活動を俯瞰



データ品質マネジメントガイドブック

AIのための高品質なデータを維持

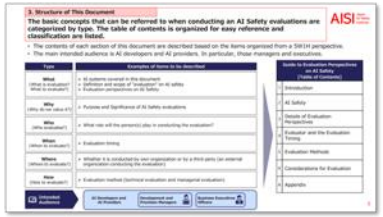


年次レポート

AISIの年間活動報告

評価観点ガイド

10の評価観点を整理



レッドチーミング手法ガイド

レッドチーミング手法の基本的な留意事項

AIセーフティ評価環境（OSSツール）

ガイドに基づく評価ツール

国際機関との関係構築

AISI Japan AI Safety Institute

国連

国連総会（UNGA80）関連行事

- 主催：United Nations Office of Digital Emerging Technology （UNODET）
- 日時：令和７年９月22日（月）9:00 – 20:00 ET
- 場所：Ease, 605 3rd Avenue, New York City
- 議題：From Principles to Practice: One Year of the GDC （オープンで安全、かつ説明責任のあるデジタルの未来の基礎となる、包括的なデジタル経済、AI ガバナンス、デジタル公共インフラ等）

OECD

日本AISI主催広島AIプロセス及びAI法関連サイドイベント

- 日時：令和7年7月18日（金）9時～12時
- 場所：University of British Columbia
- 参加：菅田参事官等（内閣）、寺村特別交渉官等（総務）、村上所長、江間顧問等（AISI）、各国安全性機関、広島AIプロセス報告枠組みの関係企業等５０名ほど
- 目的：AISI国際ネットワーク会合のサイドイベントとしてAISIが開催。安全、安心、信頼できるAIの実現に向け、広島AIプロセスの国際指針と整合した日本のAI法を各国のAI安全性機関等に説明し、国内でのAI法導入をベストプラクティスとしての普及を図る。

米国Anthropic社との協力覚書

AISI Japan AI Safety Institute

AI評価（AIモデルの能力、限界、潜在的リスク）の手法に関する情報交換とベストプラクティスの共有を推進する協力覚書

- AI評価の実効性を高めるため、AI開発企業との幅広い連携強化を推進する第一歩としてAnthropic社との協力覚書に署名（10月）
- 協力覚書に基づく活動の一環として、Anthropic社の発表したAIを悪用したサイバー攻撃に関するレポートのブリーフィングを実施（11月）
  - 中国が支援するグループによる高度なサイバー諜報活動の詳細が記載

Hiroshima Global Forum for Trustworthy AI

AISI Japan AI Safety Institute

日本のAI戦略の主要な柱の一つとしてAISIの役割と重要性を国内外に発信。Trustworthy AIの実現に向けて、国際社会やAI安全性／セキュリティ関連の産業界・研究機関等と連携し、今後のAI安全性の道筋を探ることを目的として開催。

- 日程：令和８年1月15日(木)[Open]、16日(金)[Invitation only]
- 場所：グランドプリンスホテル広島「瀬戸内」
- 参加：国内　日本AISI、連携している企業・研究機関、関係省庁等
  国外　各国AISI及びAI関係者、関連企業・研究機関等　**約150名（16か国・45名）参加**

Day1（15日）：各ステークホルダからプレゼン

- 日本のAI関連施策・今後の展望
  - AI政策/AI戦略/広島AIプロセス/AIの未来について
- AISI国際ネットワーク関係機関
  - US CAISI冒頭あいさつ
  - 各国AISIおよび関連研究機関の活動報告
  - UK AISI閉会挨拶
- 関連企業との連携
  - 日本AISIの事業実証WG
  - AIセーフティ/セキュリティに関する企業の取り組み
    - Microsoft、PFN、NTT、NEC、富士通、OpenAI、Anthropic、Cisco
- 国際社会との連携
  - OECD、GPAI Tokyo、ASEAN、United Nations University

Day2（16日）：Day1を踏まえた議論

- Roundtable1：Trustworthy AI
  - 信頼性の高いAIの課題と解決策に焦点が当てられ、透明性、安全性、相互運用性の必要性が強調された。
- Roundtable2：Secure AI
  - AIの評価を開発段階から運用段階へ拡大する点が焦点となった。
  - 主な論点として、導入前および運用段階での評価の必要性、サイバー攻撃に対するAIセキュリティの重要性、AIインシデント報告の共通フレームワーク構築が挙げられた。

AIセーフティ・インスティテュート(AISI)の機能強化について

2026-1-22
AISI事務局

AISI Japan AI Safety Institute

世界のAI安全性機関の国際ネットワーク

AISI Japan AI Safety Institute

米国の呼びかけで「International Network for AI Safety Institute」として発足
初年度の議長国は米国で現在10カ国が参加。2025年11月からは英国がコーディネーター
「The International Network for AI Evaluation and Research」に名称変更

カナダ
- 2024年11月、AISI設立

米国
- 2024年2月、NIST(国立標準技術研究所)にAISIを設立
- 2025年6月にCenter for AI Standards and Innovation(CAISI)に改名。

英国
- 2023年11月、DSIT(科学イノベーション技術省)にAISIを設立。
- 2025年2月にAI Security Instituteに改名。

EU
- 2024年5月、欧州委員会に設立されたAI OfficeがAISI相当の機能も担い、利活用に加え、安全性も推進。AI法の整備と推進も担う。

フランス
- INESIA（AISI相当機関）を2025年2月に設立

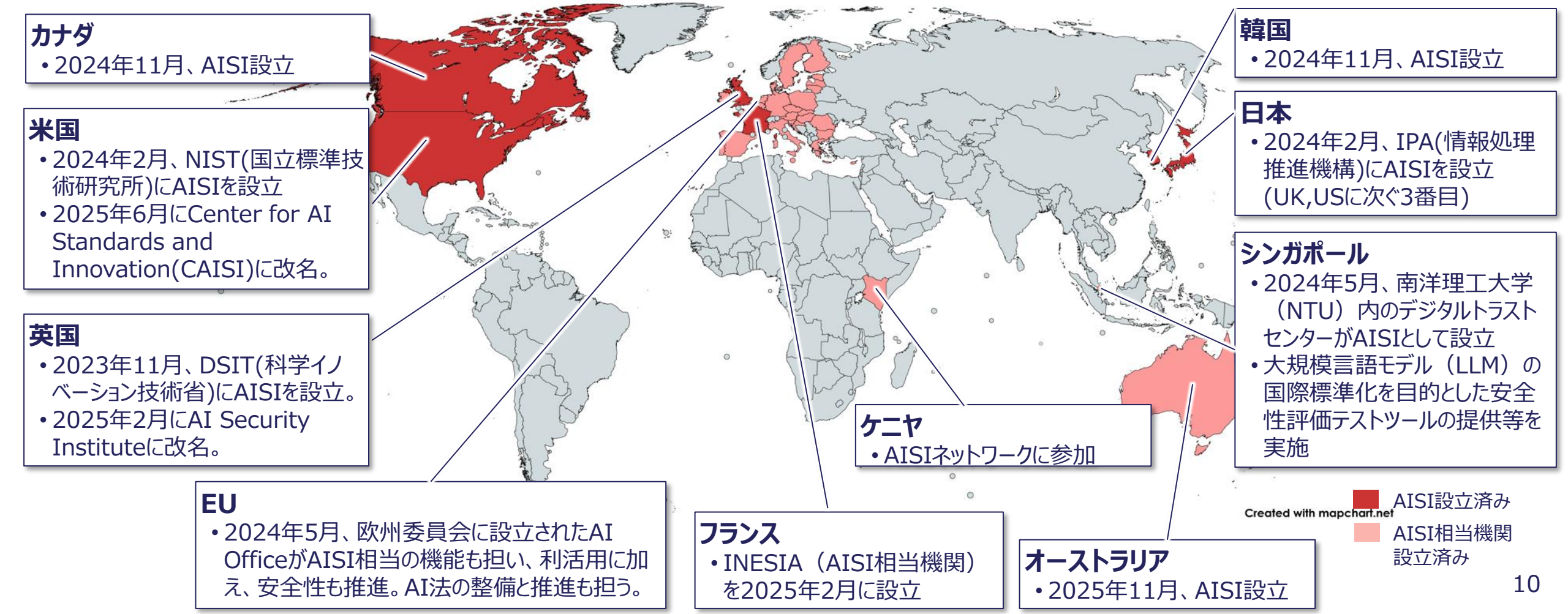
ケニヤ
- AISIネットワークに参加

オーストラリア
- 2025年11月、AISI設立

韓国
- 2024年11月、AISI設立

日本
- 2024年2月、IPA(情報処理推進機構)にAISIを設立(UK,USに次ぐ3番目)

シンガポール
- 2024年5月、南洋理工大学（NTU）内のデジタルトラストセンターがAISIとして設立
- 大規模言語モデル（LLM）の国際標準化を目的とした安全性評価テストツールの提供等を実施



3タイプある世界のAI安全保障戦略のいいとこどりを

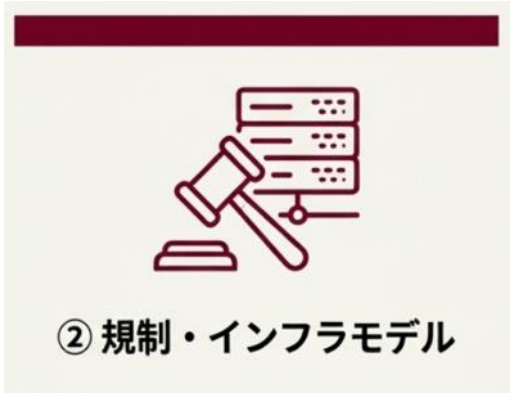
AISI Japan AI Safety Institute

「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持ちいて、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

日本AISIの方針とミッション

AISI Japan AI Safety Institute

各省・民間業界と連携し、産業分野ごとのAI安全基準を策定、必要に応じて制度化加えて、国家安全保障を含む悪用・誤作動対応を行う

通常AI安全性対応

取組方針	**基準作り**	制度化（認証）
各省所管の産業分野など、ドメインごとの課題整理や方針検討	**安全性評価のための基準（ベンチマーク）および評価手法の確立** **※データ、モデル、システム、ドメイン**	安全性評価基準に基づいた評価の実施/認証

悪用・誤作動対応

誤作動等が発生した場合の影響が大きい分野で用いられるAIの安全性の評価
影響の大きい悪用や誤作動の可能性があるAIの開発流通を防止する政策的・技術的な対応の助言

省庁・AISIと民間の役割分担

AISI Japan AI Safety Institute

官民エコシステムにより評価基準の策定を行い、制度化を支援
安全保障を含めた悪用・誤作動に技術的観点から対応する

	取組方針	基準（ベンチマーク）策定	制度化（認証の場合）		悪用・誤作動対応
関係省庁	◎	◎（政策観点）	○		○
AISI	技術観点から支援	◎（技術観点）	技術観点から支援		◎（調査・技術評価等）
民間		業界観点から参画・支援			情報提供
	各省所管の産業分野など、ドメインごとの課題整理や方針検討	安全性評価のための基準（ベンチマーク）および評価手法の確立 ※データ、モデル、システム、ドメイン	安全性評価基準に基づいた認証		誤作動等が発生した場合の影響が大きい分野で用いられるAIの安全性の評価 影響の大きい悪用や誤作動の可能性があるAIの開発流通を防止する政策的・技術的な対応の助言

AISI強化策の具体策について

AISI Japan AI Safety Institute

２つの技術領域（モデル・システムと利活用）と３つの横断施策

	モデル・システム（技術）	AI利活用
基準作り・評価手法検討	データ・学習プロセス安全確保	社会・制度と接続する技術
人材育成・啓蒙活動	モデル・アルゴリズムの安全性	悪用・濫用・デュアルユース対策
標準・国際連携	AIモデル・システムのためのセキュリティ	ドメイン特化型AIの安全性
		フィジカルAI・サイバーフィジカル安全

人材強化策（案）

AISI Japan AI Safety Institute

AISI事務局の規模拡大は必須

まずは60人、UKに遜色ないレベル（100人規模）を目指す

本日の論点

- **関係省庁からの出向強化**
  - AISI関係省庁等との調整業務
  - 諸外国の政府機関等との渉外業務
  - 予算・調達等の公務員ルール対応

現行　→ 強化後のターゲット
- 出向　3省庁→関連全省庁
- 専任20人。IPA内の併任を含めて30人強の体制→60人体制

- **国研や大学・企業のクロスアポイントメント**
  - パートナーシップ協定のさらなる強化
  - 海外在住の日本人の採用
- **民間からのの積極採用**
  - 有期雇用ではなく安定雇用へ

- AISI所属が研究者としてのキャリアにプラスとなるような仕組みが必要
- まずはパートナーシップ機関であるNII、RIKEN、AIST、NICTとの間で実現可能性を検討

AISIのR8.4月時点の組織体制の方向性（案）

AISI Japan AI Safety Institute

R7年度補正予算及び自民党提言への着実な対応のため、現在のAISI組織体制を見直し。AISIパートナーシップ機関等とも緊密に連携しつつ、業務を展開。

所長（1人）・副所長（1人＋2人）

注：早期に、党の提言にある60名体制を実現するために、省庁出向者（独法事務含む）を25名程度欲しい。主に朱色の囲みの箇所は省庁出向者のバックアップが欲しい

1．戦略・企画・全体総括

- プロジェクトマネジメント、組織運営（企画・総括、会計・契約、人事庶務、規程改正等）、渉外対応（予算要求、法改正、内閣府委託調査対応等）を実施

2．【AIの物差しを作る】 ベンチマーク、ガイドライン・基準、ツール開発等、そのために必要な個別研究・技術開発

- AI（AIエージェント、フィジカルAI含む）の安全性確保のためのガイドライン、評価ツール等開発
- そのための情報提供、AI製品・サービスのための認証制度への貢献
- これらに必要な基礎研究・技術開発（データ学習安全確保、モデル・アルゴリズムの安全確保、フロンティアモデルAI安全、社会と制度を接続する技術・ガバナンス技術）

3．【AI政策を実装する、AIを運用する】 影響の大きなAIの安全性評価、それを踏まえた対策の助言

- AIに係るサイバー攻撃・AIによる防御、
- AIによる悪用・濫用、デュアルユース対策、
- 国の基盤となる分野におけるAI安全性評価（事業実証WG、コンソーシアム運営等）
- 影響の大きなAIの開発・流通を防止する仕組みづくり（民間企業等の開発したAIの評価含む）

【国の基盤となる分野（事業実証WG）】

ヘルスケア　ロボティクス　（仮）教育　（仮）金融　…

4．【国際協調する、実態を把握する】 国際連携・協調、AI利用実態調査

- ２．３の取組の海外AISI機関等との連携・協力、
- 不適切なAIの開発・利用の事案把握、情報提供（性的被害、差別的被害対応等）

AISI

Japan AI Safety Institute