AIセーフティ・インスティテュート（AISI）について

※AISIは、エイシーと読みます。

2025-04-01
AISI事務局

AISI Japan AI Safety Institute

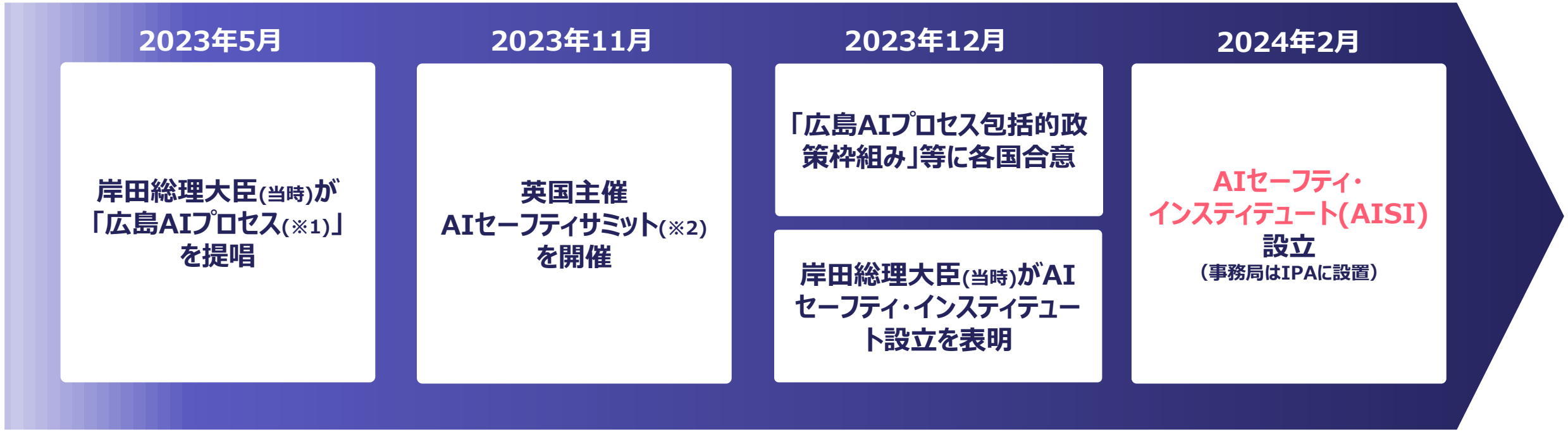
AISI Japan AI Safety Institute

AISI設立の背景・概要

日本におけるAISIの設立

AISI Japan AI Safety Institute

広島AIプロセスでの議論や**AIセーフティサミット**を経て
日本でもAIセーフティ・インスティテュート（AISI）を設立（2024年2月）



※1 成果文書｜広島AIプロセス
※2 AI Safety Summit 2023 - GOV.UK

統合イノベーション戦略2024

AISI Japan AI Safety Institute

「統合イノベーション戦略2024」において、AISIは**日本におけるAIの安全性の中心機関**と定義

- 統合イノベーション戦略2024とは、内閣府による第6期科学技術・イノベーション基本計画の実行計画として位置付けられる４年目の年次戦略であり、３つの強化方策を打ち出すとともに、従来からの3つの基軸についても着実に推進することとしている。

統合イノベーション戦略2024における３つの強化方策

(1) 重要技術に関する統合的な戦略

(2) グローバルな視点での連携強化

(3) AI分野の競争力強化と安全・安心の確保

①AIのイノベーションとAIによるイノベーションの加速（研究開発力の強化、AI利活用の推進、インフラの高度化等）
②AIの安全・安心の確保（ガバナンス、AIの安全性の検討、偽・誤情報への対策、知財等）
③国際的な連携・協調の推進（広島AIプロセスの成果を踏まえた国際連携等）

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

- AISIは自ら研究開発を行う組織ではない

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

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**（英米のAISI等）**との国際連携に関する業務**

AISI Japan AI Safety Institute

体制

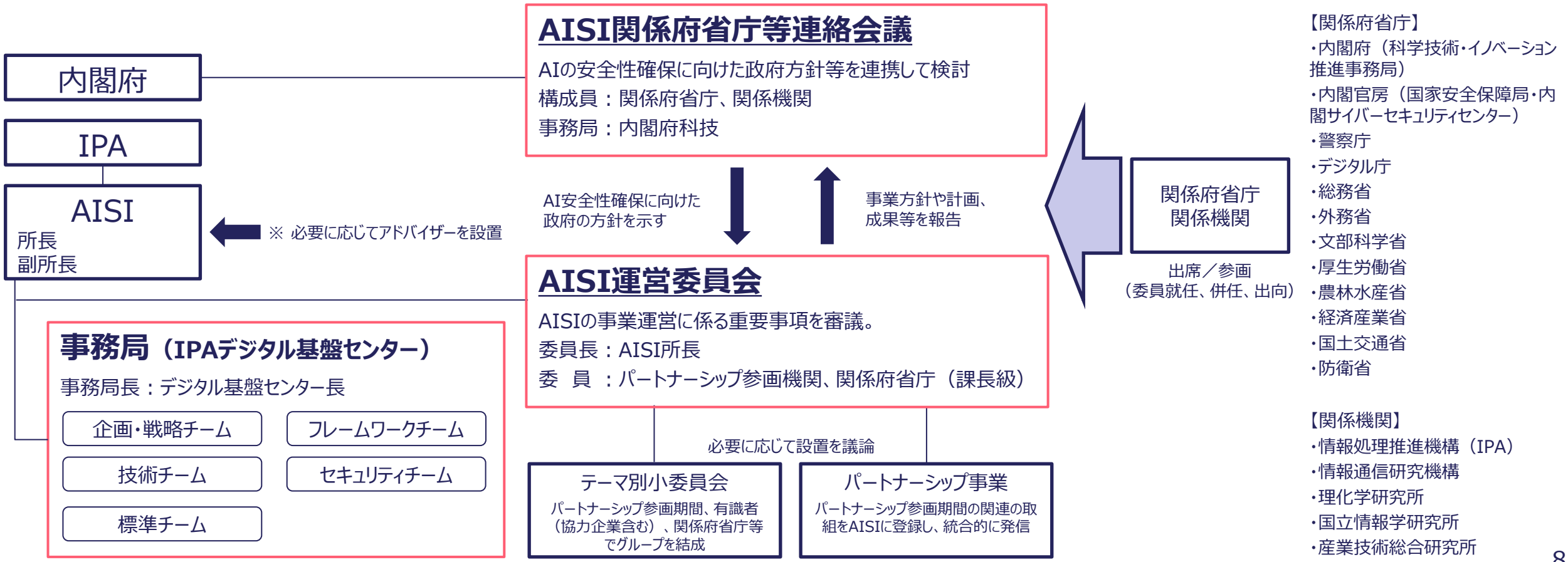
AISIの推進体制

AISI Japan AI Safety Institute

AISIは13府省庁、5関係機関で構成される**政府横断の組織**

＊内閣府を事務局とする「AISI関係府省庁等連絡会議」で政府方針等を検討

＊AISIには「AISI運営委員会」と「事務局」を設置



AISIパートナーシップ協定の発効

AISI Japan AI Safety Institute

関係府省庁の協力の下、関係機関が連携してAIの安全性に係る取組を推進していく協力体制（AISIパートナーシップ協定）を発行（2024年8月）

- AIの安全性に関する取組を進めるためには、**AISIのみならず、国内の関係機関と連携し、共同で対応していく**ことが不可欠。

- 以下、「日本AIセーフティ・インスティテュートパートナーシップ協定」より一部抜粋

第３条　（パートナーシップの活動内容）

日本AIセーフティ・インスティテュートパートナーシップ（以下「本パートナーシップ」という。）は、第５条の規定に基づく参画機関との協力の下、AISIの活動を効果的に推進するため、第７条第２項の規定に基づきAISIと参画機関との間で合意した範囲において、次の活動を推進する。

① AI安全性に関してAISIと参画機関が共同で実施する研究及び調査
② AISIが実施する活動に関する参画機関による助言の付与
③ 参画機関が実施するAI安全性に関する活動についてのAISIへの情報提供
④ 前各号の取組に関するAISI及び参画機関による国内外への情報発信、国内外の関係機関との調整・連携
⑤ その他前各号の活動に附帯する活動

AISIの幹部メンバー

AISI Japan AI Safety Institute

所長　**村上　明子**

1999年　4月	日本アイ･ビー･エム株式会社 東京基礎研究所 入社
2016年　1月	同社 東京ソフトウェア開発研究所
2021年　4月	損害保険ジャパン株式会社 入社 執行役員待遇 DX推進部 特命部長
2021年10月	同社 執行役員待遇 DX推進部長
2022年　4月	同社 執行役員 CDO(Chief Digital Officer)　DX推進部長
2024年　4月	同社 執行役員 CDaO(Chief Data Officer)　データドリブン経営推進部長　［現職兼務］
2025年　4月	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［現職兼務］

副所長･事務局長

平本　健二

1990年4月	NTTデータ通信株式会社 入社（現 株式会社NTTデータ）
2008年7月	経済産業省 CIO補佐官
2012年8月	内閣官房 政府CIO上席補佐官
2021年9月	デジタル庁 データ戦略統括
2023年7月	IPAデジタル基盤センター センター長　［現職兼務］
2024年2月	AISI事務局長（4月より副所長兼務）

副所長

寺岡　秀札

1999年4月	郵政省　入省
2007年7月	総務省総合通信基盤局
2023年7月	内閣官房内閣サイバーセキュリティセンター
2024年4月	AISI副所長

AISI 事務局

AISI Japan AI Safety Institute

AISI事務局は、以下５つのチームで構成されており、
政府や民間企業からの出向者も多数在籍

戦略・企画

- AISIの戦略や計画の作成、予算管理
- 広報、採用、人材育成
- 国内外の関係機関との調整・支援

技術

- AIセーフティに関する評価方法の確立
- 評価環境の開発

標準

- AI分野における適合性評価の手法確立
- 実運用を見越した国内体制構築の検討

フレームワーク

- AIセーフティに関する評価の枠組みの検討
- AIガバナンスに関する相互運用性確保に向けた調整

セキュリティ

- AIシステムに対する特有の攻撃手法の調査
- AIセキュリティインシデントの分類体系の検討
- AIシステムを狙った攻撃を体系化

取組・成果物

2024年度の活動と成果物

AISI Japan AI Safety Institute

		国際	AISI	政府
		イベント	成果物	
2024	4月		• **日米クロスウォーク1**の成果公表(4/30)	• **AI事業者ガイドライン**の公表(4/19)
	5月	AIソウル・サミット, 韓国		
	6月	G7サミット, イタリア		• **統合イノベーション戦略2024**の公表(6/4)
	7月		• 米国AI RMF 日本語翻訳版の公開(7/4)	
	8月		• **評価観点ガイド**の公表(9/18) • **日米クロスウォーク2**の成果公表(9/18) • **レッドチーミング手法ガイド**※の公表(9/25)	
	9月			
	10月			
	11月	AISI国際ネットワーク会合, 米国		
	12月			
2025	1月		• **AIセーフティに関する活動マップ**の公表(2/7) • **データ品質マネジメントガイドブック**(ドラフト版)の公開(2/7) • **年次レポート**の公表(2/5)	
	2月	AIアクションサミット, フランス		
	3月			• **AI事業者ガイドライン**の更新(3/28)

※レッドチーミングとは、攻撃者の目線で対象AIシステムにおける弱点や対策の不備を発見し、これらを修正及び堅牢化することで、AIセーフティを維持または向上させる取り組み

成果物の概要

AISI Japan AI Safety Institute

AI事業者ガイドラインを軸に、
技術的なレビューから人材育成まで、幅広く取り組む

クロスウォーク
国際的な相互運用性のため

AI事業者ガイドライン

活動マップ
全体像と優先順位付け

評価観点ガイド
評価

レッドチーミング手法ガイド
レッドチーミング

データ品質マネジメントガイドブック
AIに適格なデータを提供するため

多言語/多文化
多国間での問題

セキュリティレポート
セキュリティに関する知識

デジタルスキル標準
人材育成

AI事業者ガイドラインの概要

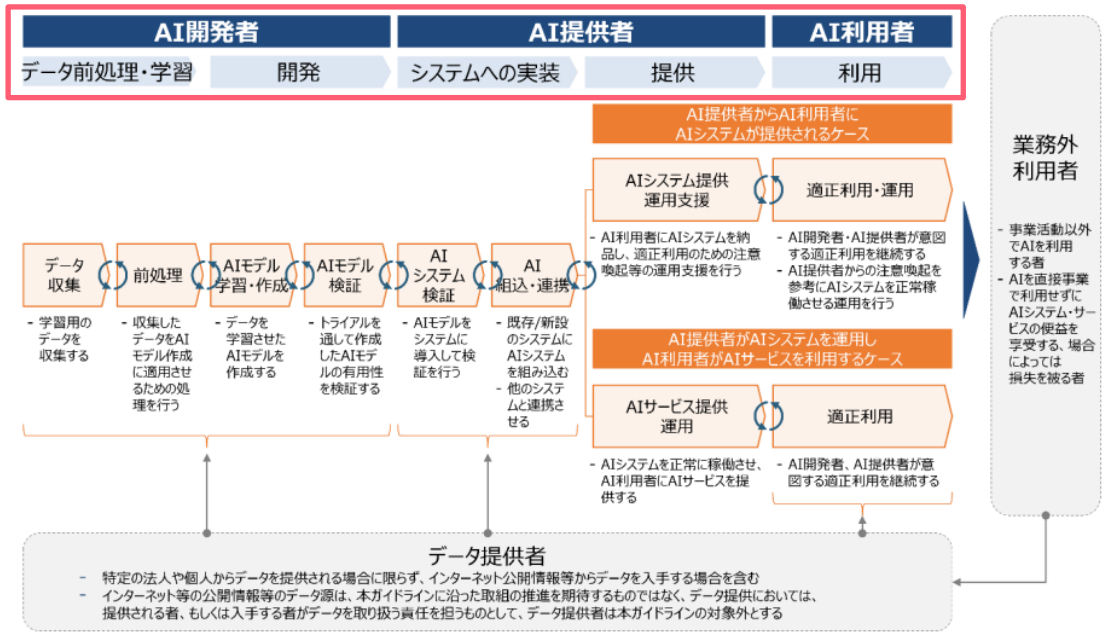
AISI Japan AI Safety Institute

経済産業省及び総務省は、既存のガイドラインを統合・アップデートし、「AI事業者ガイドライン（第1.0版）」を公表 ※2025年3月 1.1版に更新

- AI活用の流れの中で、各ステークホルダが対応すべきことを明確化。
- AISIは、AI事業者ガイドライン検討会を経済産業省と共同事務局として開催、運営している。

事業者のガイド利用を意識した内容

AI活用ライフサイクル・主体毎に対応すべき内容を定義



民間事業者によるガバナンスの行動目標※を定義

※「3-1-1」のように更に細分化されているものもあり

分類	行動目標
1. 環境・リスク分析	1-1 便益/リスクの理解 1-2 AIの社会的な受容の理解 1-3 自社のAI習熟度の理解
2. ゴール設定	2-1 AIガバナンス・ゴールの設定
3. システムデザイン	3-1 ゴールと乖離の評価及び乖離対応の必須化 3-2 AIマネジメントの人材のリテラシー向上 3-3 各主体間・部門間の協力によるAIマネジメント強化 3-4 予防・早期対応による利用者のインシデント関連の負担軽減
4. 運用	4-1 AIマネジメントシステム運用状況の説明可能な状態の確保 4-2 個々のAIシステム運用状況の説明可能な状態の確保 4-3 AIガバナンスの実践状況の積極的な開示の検討
5. 評価	5-1 AIマネジメントシステムの機能の検証 5-2 社外ステークホルダーの意見の検討
6. 環境・リスクの再分析	6-1 行動目標1-1～1-3の適時の再実施

※ 中小企業においても、まずはどのようなアクションをとればよいのか明確化している

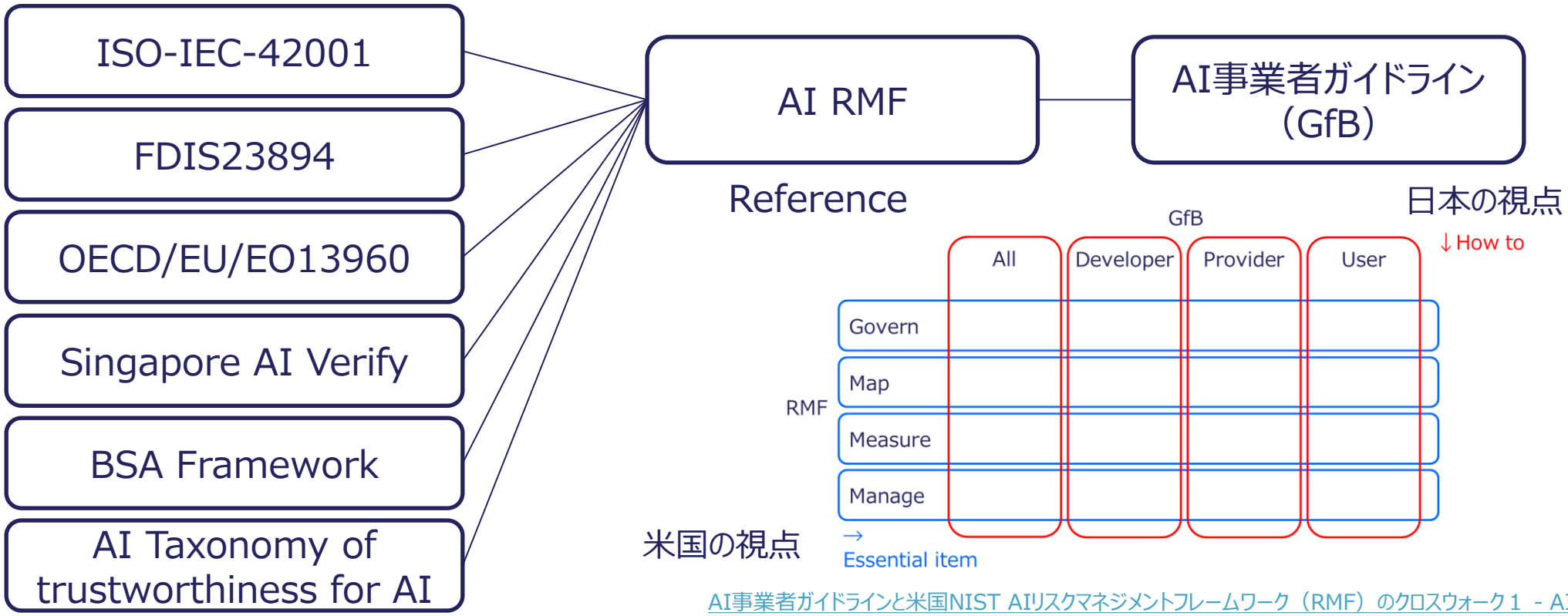
出典）「AI事業者ガイドライン（第1.1版）」本編（概要）及び別添（概要）総務省・経済産業省 より抜粋

日米クロスウォークの概要

AISI Japan AI Safety Institute

米国NISTのAI Risk Management Framework(RMF)と日本のAI事業者ガイドライン(Guidelines for Business; GfB)の相互関係を確認

- 米国のAI RMFをリファレンスに各国ガイドライン等との確認も可能。



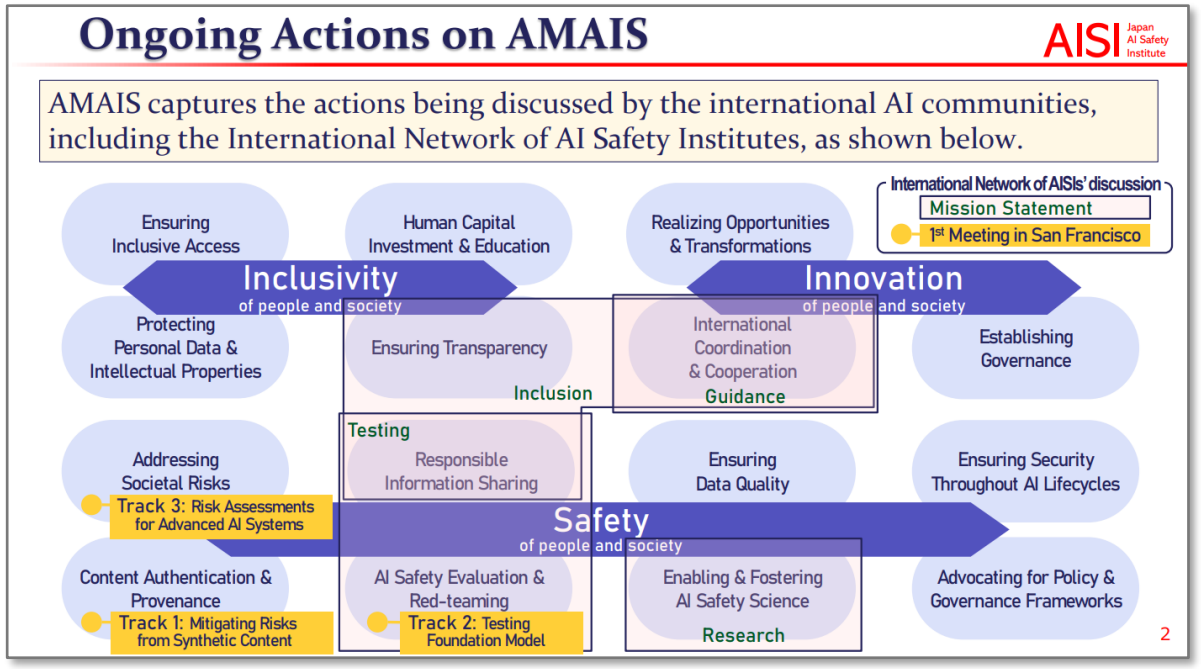
AI事業者ガイドラインと米国NIST AIリスクマネジメントフレームワーク（RMF）のクロスウォーク１ - AISI Japan
AI事業者ガイドラインと米国NIST AIリスクマネジメントフレームワーク（RMF）のクロスウォーク２ - AISI Japan

AIセーフティに関する活動マップ（AMAIS）の概要

AIの安全性に関する活動が急速に変化・進化する中、
見落とされがちな部分や活動間の相関関係を全体像として可視化

- AISIは、**ディスカッションペーパー**として「AMAIS：AIの安全性に関する活動マップ」を公開。

- AISIは、**主要文献のベンチマーク**に基づき、包括的なアクティビティマップと関連用語を開発している。
- この日本主導の取り組みは、AIの安全性に関する国際的な協力体制の基盤をさらに強化し、持続可能で信頼性の高いAI社会の実現に貢献することが期待されている。



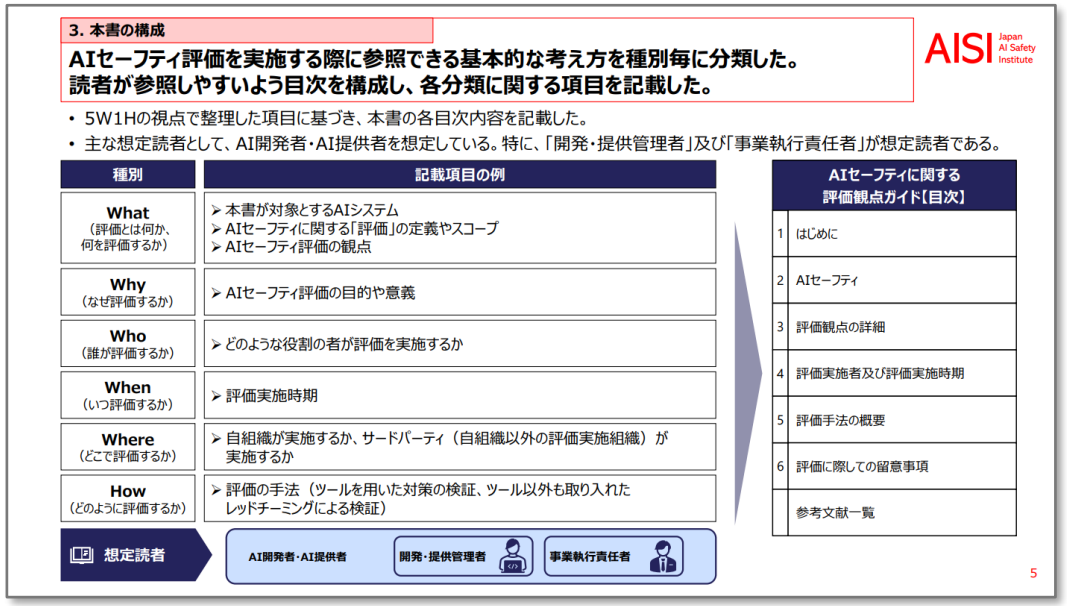
AIセーフティに関する活動マップ（AMAIS）AMAIS: Activity Map on AI Safety - AISI Japan

評価観点ガイドの概要

事業者がAIを開発・提供する際の参考として、
AIシステムの安全性を評価する際の基本的な考え方を示したもの

- 具体的には、以下の事項等が記載されている。
  - **安全性評価で想定するリスクや評価項目**
  - **評価の実施者や実施時期**
  - **評価手法の概要**

- このガイドは、安全・安心で信頼できるAIの実現に向けての第一歩であり、今後のAI開発・提供における安全性の維持・向上に資することを期待している。

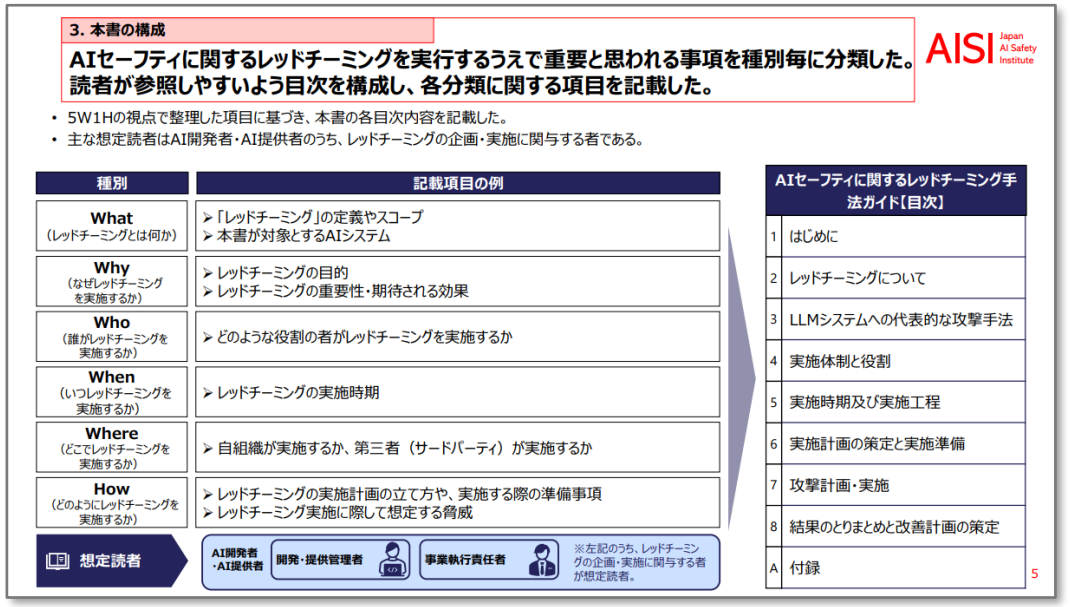


AIセーフティに関する評価観点ガイド - AISI Japan

レッドチーミング手法ガイドの概要

事業者が開発・提供する際の参考として、**AIシステムの安全性を評価する手法**の１つであるレッドチーミング手法について基本的な留意事項を示したもの

- 具体的には、安全性評価の実施体制、時期、計画、実施方法、改善計画の策定等にあたっての留意点が示されている。
- このガイドは、安全・安心で信頼できるAIの実現に向けての第一歩であり、今後のAI開発・提供における安全性の維持・向上に資することを期待している。

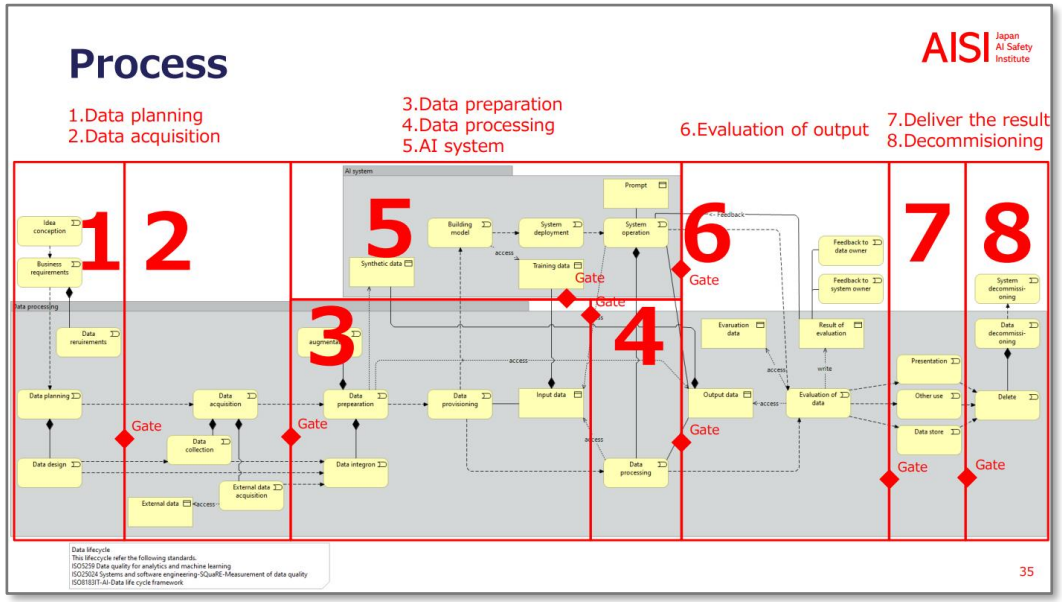


AIセーフティに関するレッドチーミング手法ガイド - AISI Japan

データ品質マネジメントガイドブックの概要

データとAIの価値を最大化するために必要な
データ品質を持続的に確保するため、何をすべきか整理

- データ品質は、AIの卓越性の基礎であり、信頼できるAIの実現に寄与する。AI社会を適切に実現し、データ駆動型社会へと導くため、本ガイドに整理。
- 本ガイドは、英語版が正式版であり、2025年3月に日本語訳サマリが公開。

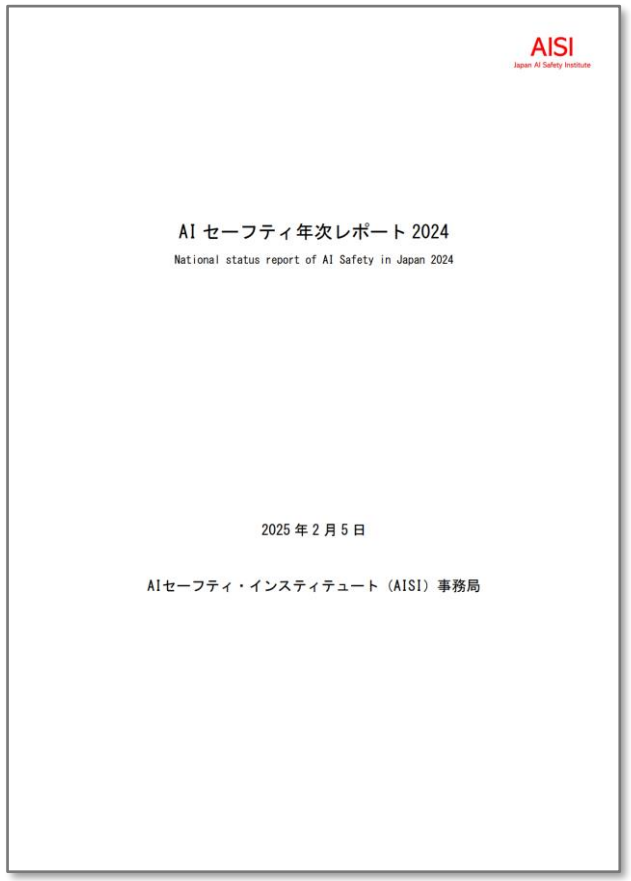


AIセーフティ年次レポート2024の概要

AISIの活動状況を「AIセーフティ年次レポート2024」としてまとめた。

- 「AIセーフティ年次レポート2024」とともに、関連するレポート等についても、年次レポートを補完する参考資料「AIセーフティ　ファクトシート2024」として取りまとめた。

- 本稿においては、AIの急速な進展に対応するための、AISIと国内外の関係機関や企業等との連携など、我々の今後の取り組みやその狙いについても記載している。



AIセーフティ年次レポート2024 National Status Report on AI Safety in Japan 2024 - AISI Japan

AISI Japan AI Safety Institute

国際連携

主要な国際会合

AISI Japan AI Safety Institute

AIセーフティ関連の国際会合に積極的に参加するとともに、
各国のAI関連事業者及び団体との意見交換も実施

- AISI関連のトップレベルの連携
  - **スタンフォード大学AIシンポジウム（2024年4月16日、スタンフォード）**
    - 米国・英国AISIの所長等とパネルディスカッション、並行した各国間意見交換
  - **AIソウル・サミット（2024年5月21-22日、ソウル）**
    - ハイレベルラウンドテーブル他、米英EU加独などと意見交換
    - 同時開催のAIグローバルフォーラムでアジア、アフリカ諸国等を含む議論に参加
  - **国連未来サミット（2024年9月22日、国連本部）**
  - **国連Global Compact Leaders Summit 2024（2024年9月24日、国連本部）**
    - 各国AI責任者などとAIセーフティに関して議論
  - **AISI国際ネットワーク会合（2024年11月10-11日、サンフランシスコ）**
  - **AIアクションサミット（2025年2月6-11日、パリ）**
  - **広島AIプロセス・フレンズグループ会合（2025年2月27-28日、東京）**

AIソウルサミット同時開催の
グローバルフォーラム


国連未来サミット

AISI国際ネットワークメンバーの組織概要

AISI Japan AI Safety Institute

各国AISIの設立状況については、以下の通り（2024年10月のCSISによるレポート）。

	米国	英国	EU	日本	シンガポール	韓国	カナダ
設立	2024年 2月	2023年 11月	2024年 5月	2024年 2月	2024年 5月	2024年 5月 (発表)	2024年 4月 (発表)
名称	US AISI	UK AISI	EU AI Office	Japan AISI	Singapore AISI	Korea AISI	Canada AISI
親組織	National Institute of Standards &Technology (NIST)	Department for Science, Innovation &Technology (DSIT)	Directorate-General for Communications Networks, Content & Technology (CNECT)	独立行政法人 情報処理推進機構 (IPA)	Digital Trust Centre (DTC)	Electronics & Telecommunications Research Institute (ETRI)	
予算	1,000万ドル (FY24)	6,500万ドル (年間)	5,100万ドル (期間不明)		750万ドル (年間)	720万~ 1,440万ドル (年間)	3,650万ドル (期間不明)
人員	20名 (コアスタッフ)	20名 (コアスタッフ)	50名 (AIセーフティユニットの計画値)	23名		最低30名 (計画値)	
機能	AISIの ビジョン・ミッション・ 戦略目標	AISIの紹介	AI Officeのタスク	AISIのタスク	初期の研究分野		
研究・ ガイドライン	デュアルユース基盤モデルの誤用リスク管理	ウェブサイトを参照		ウェブサイトを参照	生成AIのための AIガバナンスのフレームワークモデル		

公開情報

公開情報なし

出典）The AI Safety Institute International Network, Next Steps and Recommendations, By Gregory C.Allen and Georgia Adamson (P.11)の図を一部改変

AISI

Japan AI Safety Institute