

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。

2025-05-01
AISI事務局

AISI設立の背景・概要

日本におけるAISIの設立

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）

2023年5月

岸田総理大臣(当時)が
「広島AIプロセス(※1)」
を提唱

2023年11月

英国主催
AIセーフティサミット(※2)
を開催

2023年12月

「広島AIプロセス包括的政
策枠組み」等に各国合意

岸田総理大臣(当時)がAI
セーフティ・インスティテュー
ト設立を表明

2024年2月

AIセーフティ・
インスティテュート(AISI)
設立
(事務局はIPAに設置)

※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」において、 AISII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

- ① AI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）
- ② **AI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）

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AISII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

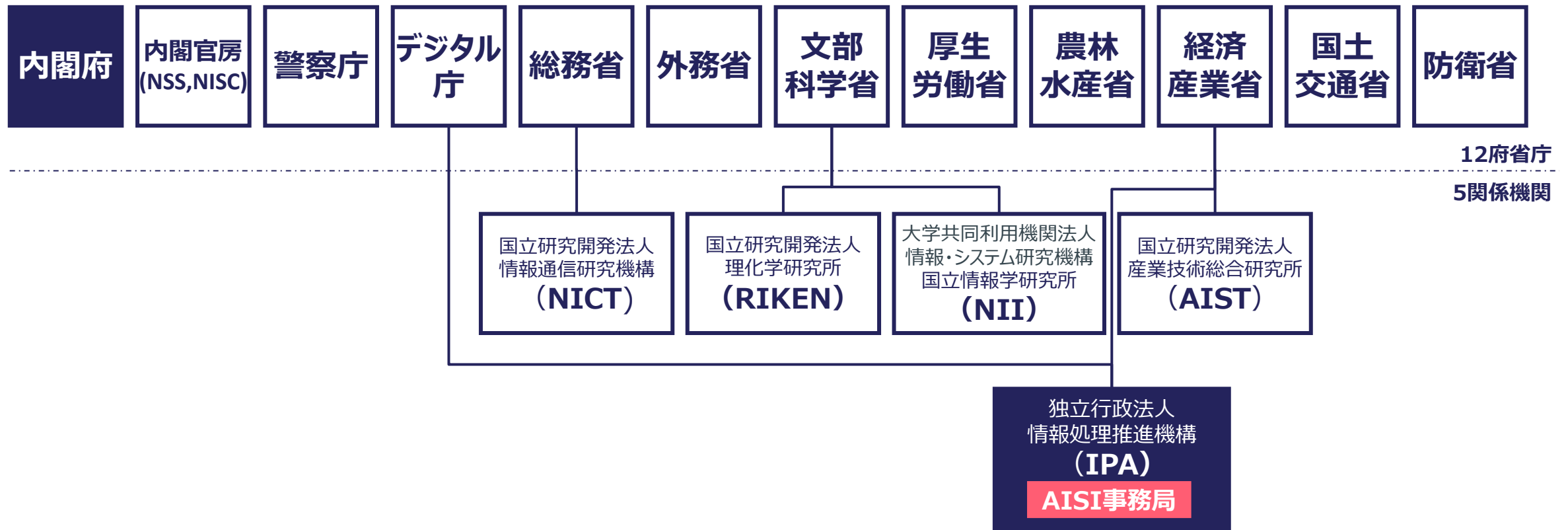
3. 他国の関係機関（英米のAISII等）との国際連携に関する業務

体制

AISIの関係府省庁・機関

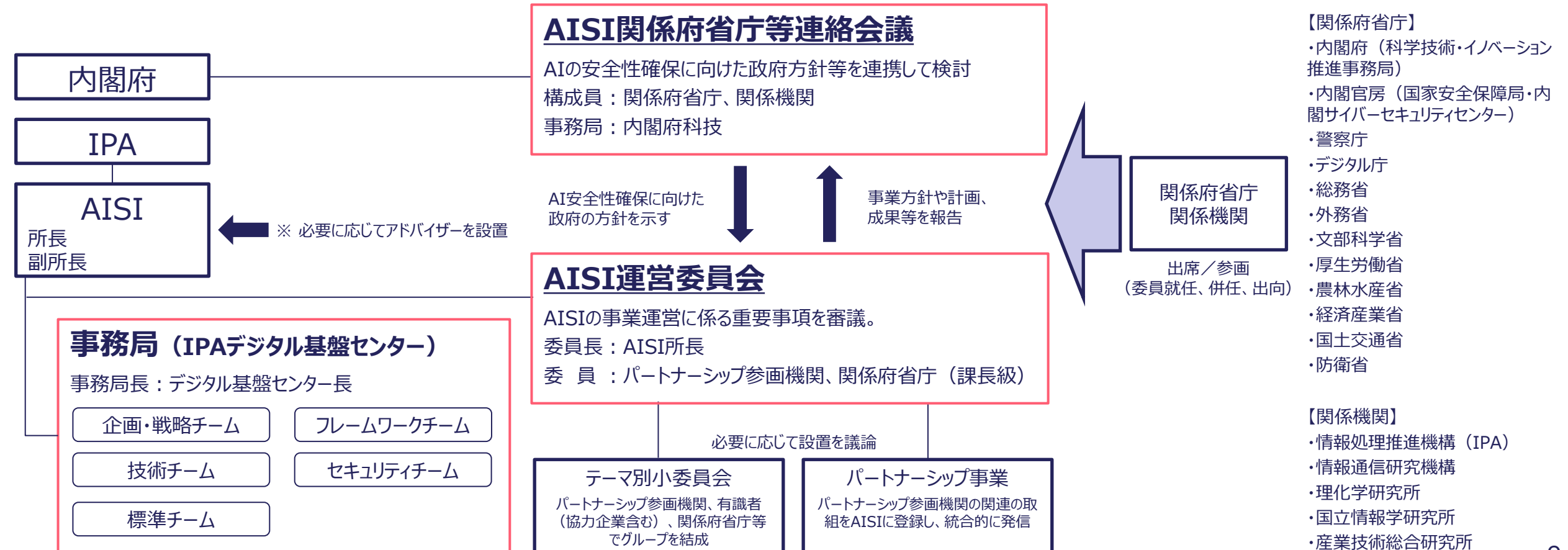
AISIは、12府省庁・5関係機関が横断的に参画する**政府関係機関**
事務局は経済産業省とデジタル庁を所管官庁としている**IPA内**に設置

* 2025年4月時点



AISIの推進体制

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所長を委員長とする「**AISI運営委員会**」で事業方針を検討



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（AISIPartner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、AISIPartnerシップ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AISIPartnerシップ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AISIPartnerシップと参画機関が共同で実施する研究及び調査
- ② AISIPartnerシップ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AISIPartnerシップへの情報提供
- ④ 前各号の取組に関するAISIPartnerシップ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

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

		国際	AISI	政府
		イベント	成果物	
2024	4月		<ul style="list-style-type: none"> 日米クロスウォーク1の成果公表(4/30) 	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 米国AI RMF 日本語翻訳版の公表(7/4) 	<ul style="list-style-type: none"> 統合イノベーション戦略2024の公表(6/4)
	7月			
	8月		<ul style="list-style-type: none"> 評価観点ガイドの公表(9/18) 	
	9月		<ul style="list-style-type: none"> 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) 	
	10月			
	11月	AISI国際ネットワーク会合, 米国	<ul style="list-style-type: none"> 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)の公表(2/7) 	
	12月		<ul style="list-style-type: none"> 年次レポートの公表(2/5) セキュリティ攻撃の俯瞰図の公表(3/26) 	
	2025	1月		<ul style="list-style-type: none"> AIセーフティの普及に向けた文書の公表(3/26)
2月		AIアクションサミット, フランス	<ul style="list-style-type: none"> 評価観点ガイド1.10版の公表(3/28) レッドチーミング手法ガイド1.10版の公表(3/31) データ品質マネジメントガイドブックの公表(3/31) 	<ul style="list-style-type: none"> AI制度研究会・中間とりまとめの公表(2/4) AI事業者ガイドライン1.1版の公表(3/28)
3月				

※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

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

総務省・経産省が策定・更新

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

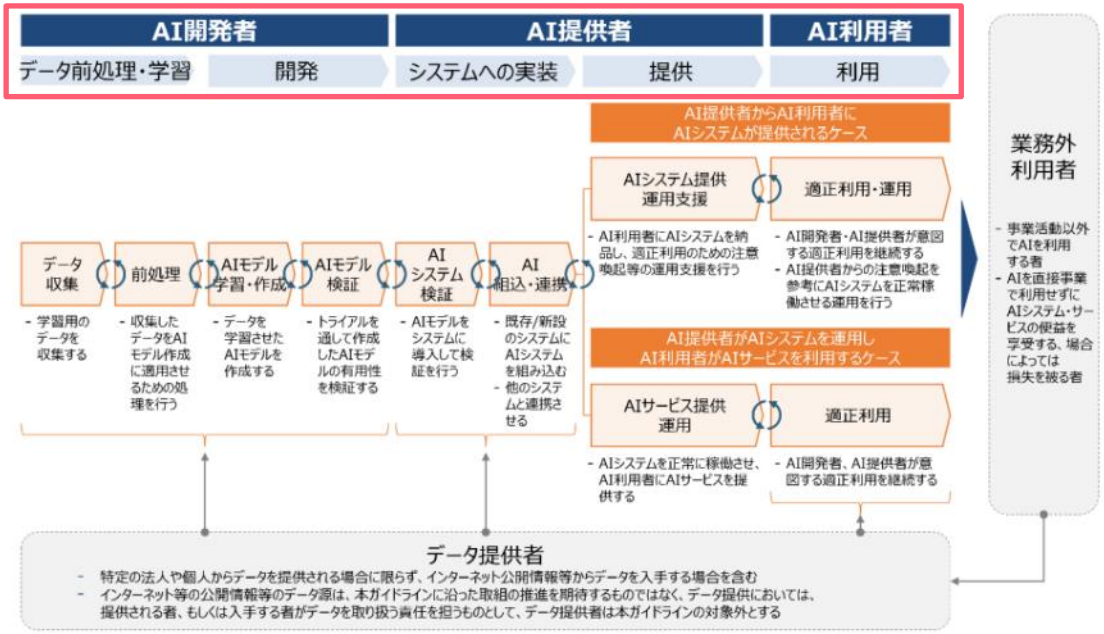
AI事業者ガイドラインの概要

総務省及び経済産業省は、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

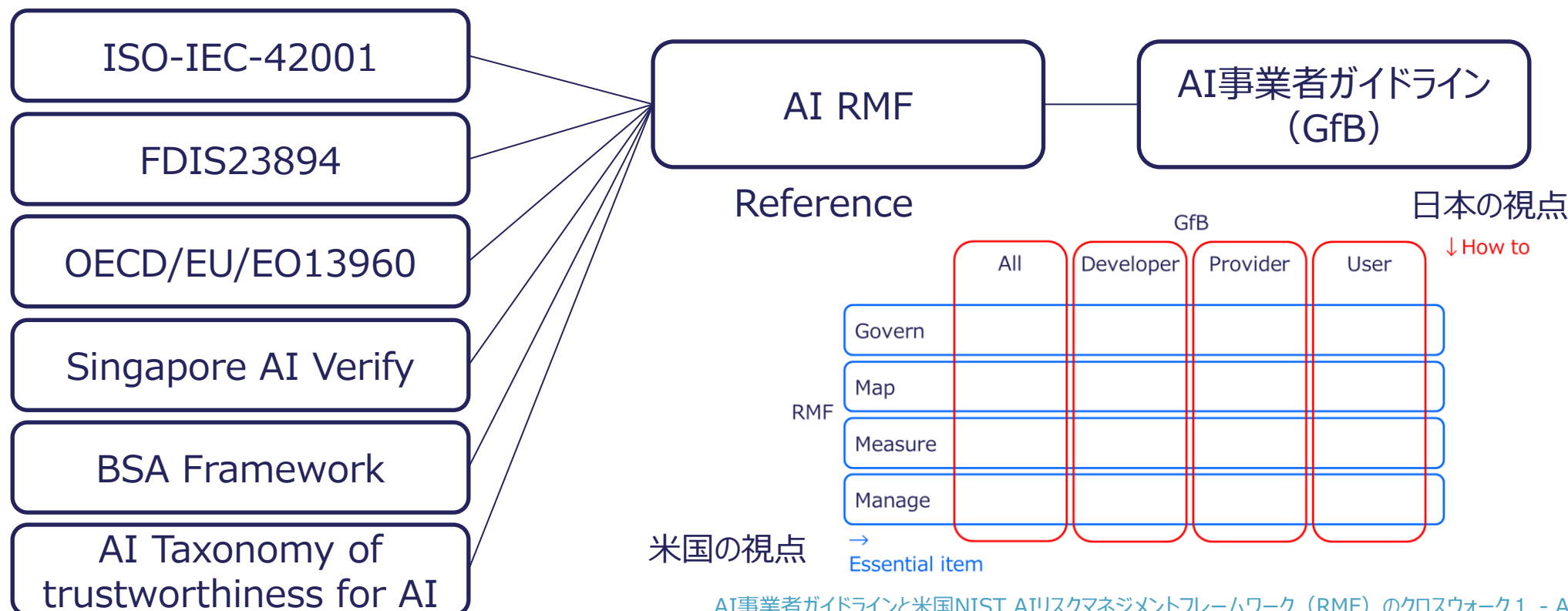


民間事業者によるガバナンスの行動目標※を定義

分類	行動目標	※「3-1-1」のように更に細分化されているものもあり
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	※ 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か明確化している
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	

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

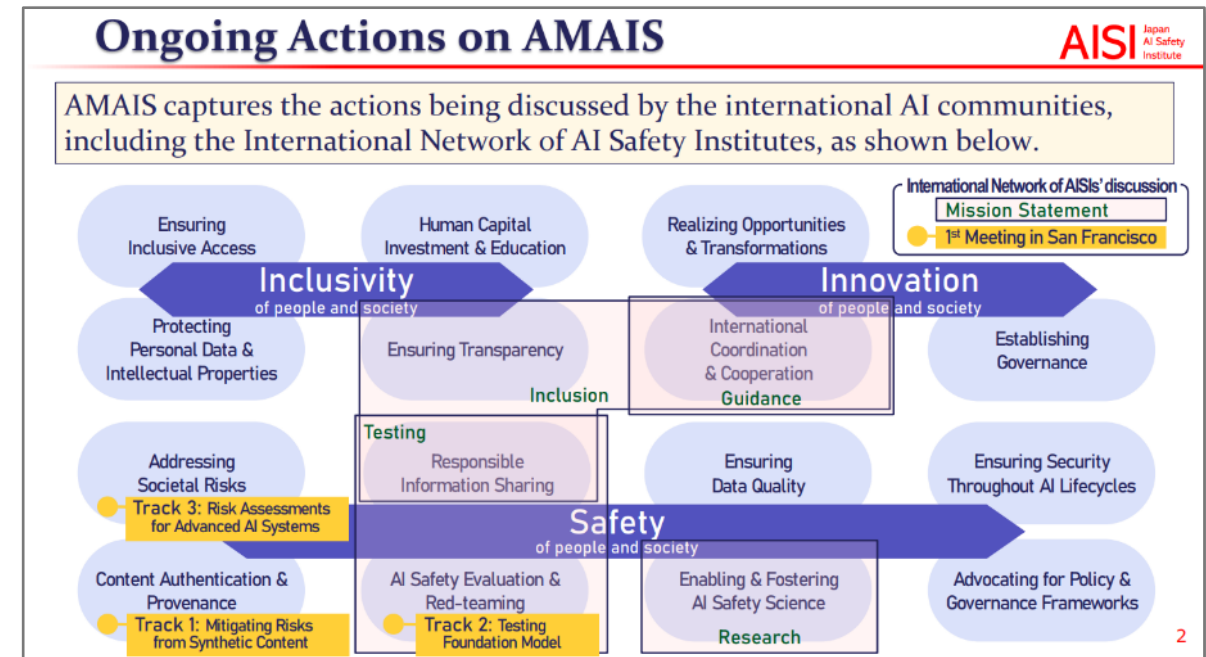
- ◆ 米国のAI RMFをリファレンスに各国ガイドライン等との確認も可能。



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。



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。

3. 本書の構成

AIセーフティ評価を実施する際に参照できる基本的な考え方を種別毎に分類した。読者が参照しやすいよう目次を構成し、各分類に関する項目を記載した。

- 5W1Hの視点で整理した項目に基づき、本書の各目次内容を記載した。
- 主な想定読者として、AI開発者・AI提供者を想定している。特に、「開発・提供管理者」及び「事業執行責任者」が想定読者である。

種別	記載項目の例
What (評価とは何か、何を評価するか)	➢ 本書が対象とするAIシステム ➢ AIセーフティに関する「評価」の定義やスコープ ➢ AIセーフティ評価の観点
Why (なぜ評価するか)	➢ AIセーフティ評価の目的や意義
Who (誰が評価するか)	➢ どのような役割の者が評価を実施するか
When (いつ評価するか)	➢ 評価実施時期
Where (どこで評価するか)	➢ 自組織が実施するか、サードパーティ（自組織以外の評価実施組織）が実施するか
How (どのように評価するか)	➢ 評価の手法（ツールを用いた対策の検証、ツール以外も取り入れたレッドチーミングによる検証）

想定読者

AI開発者・AI提供者 開発・提供管理者 事業執行責任者

AIセーフティに関する評価観点ガイド【目次】	
1	はじめに
2	AIセーフティ
3	評価観点の詳細
4	評価実施者及び評価実施時期
5	評価手法の概要
6	評価に際しての留意事項
	参考文献一覧

5

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。

3. 本書の構成

AIセーフティに関するレッドチーミングを実行するうえで重要と思われる事項を種別毎に分類した。読者が参照しやすいよう目次を構成し、各分類に関する項目を記載した。

- 5W1Hの視点で整理した項目に基づき、本書の各目次内容を記載した。
- 主な想定読者はAI開発者・AI提供者のうち、レッドチーミングの企画・実施に関与する者である。

種別	記載項目の例
What (レッドチーミングとは何か)	▶ 「レッドチーミング」の定義やスコープ ▶ 本書が対象とするAIシステム
Why (なぜレッドチーミングを実施するか)	▶ レッドチーミングの目的 ▶ レッドチーミングの重要性・期待される効果
Who (誰がレッドチーミングを実施するか)	▶ どのような役割の者がレッドチーミングを実施するか
When (いつレッドチーミングを実施するか)	▶ レッドチーミングの実施時期
Where (どこでレッドチーミングを実施するか)	▶ 自組織が実施するか、第三者（サードパーティ）が実施するか
How (どのようにレッドチーミングを実施するか)	▶ レッドチーミングの実施計画の立て方や、実施する際の準備事項 ▶ レッドチーミング実施に際して想定する脅威

想定読者

AI開発者 AI提供者 開発・提供管理者 事業執行責任者

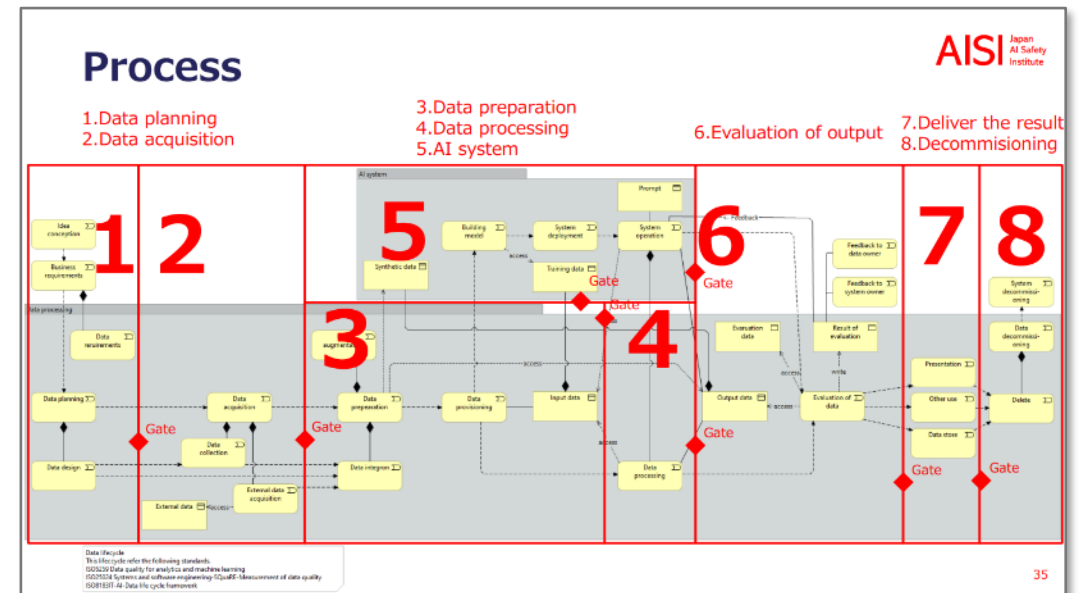
※左記のうち、レッドチーミングの企画・実施に関与する者が想定読者。

AIセーフティに関するレッドチーミング手法ガイド【目次】	
1	はじめに
2	レッドチーミングについて
3	LLMシステムへの代表的な攻撃手法
4	実施体制と役割
5	実施時期及び実施工程
6	実施計画の策定と実施準備
7	攻撃計画・実施
8	結果のとりまとめと改善計画の策定
A	付録

データ品質マネジメントガイドブックの概要

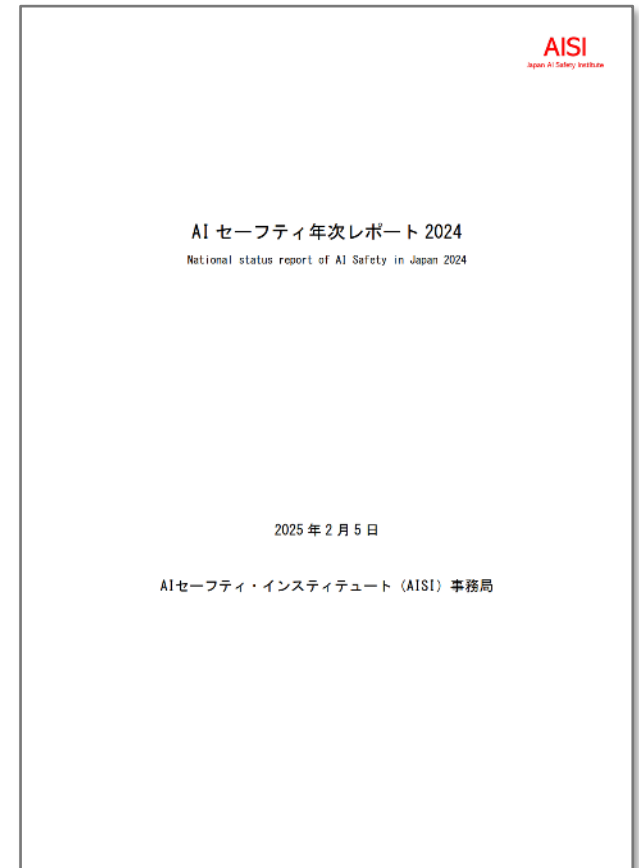
データとAIの価値を最大化するために必要な データ品質を持続的に確保するため、何をすべきか整理

- ◆ 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。



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、AISII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。



国際連携

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AISII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

米、英、EU、日、星、豪、加、仏、韓、ケニア（10カ国・地域）で構成

※各国AISII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	ウェブサイト参照		ウェブサイト参照	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

AISI

Japan AI Safety Institute