

AISIの活動

2026-10-07

全体の流れ

2023-10 

国連AIアドバイザリーボード
(HLAB)

2024-09 

UN Global Digital Compact
(GDC)

2025-08 

国連AIに関する独立国際科学パネル(IISP-AI)

2019-05 

OECD AI原則

2024-05 

OECD AI 原則(rev.)

2020-06 

Global Partnership of AI
(GPAI)

2024-07 

GPAI Tokyo Expert Support Center

2023-05 

広島AIプロセス
(HAIP)

2023-10 

HAIP行動指針
& 行動規約

2024-05 

HAIPフレンズグループ

2025-02 

HAIP報告枠組み (レポーティングフレームワーク)
(HAIP RF)

2026-06 

HAIP RF(v2.)

2024-02
AISI設立



2025-05 

AI法
(9月施行)

2025-12 

AI基本計画

2026-07 

AI基本計画2、
バーティカルAI戦略

2026-05 

プロジェクトYATA-Shield 2

2018 -> 

AI戦略会議

2023-05 -> 

AI戦略本部

— スコープ・対応分野の広がり

技術領域、適用業務領域、検討範囲など、スコープが広がっている

技術領域	適用業務領域	検討範囲	検討対象
• フロンティアAI • エージェントAI • マルチモーダル • 合成コンテンツ、	• ヘルスケア • ロボット • 教育 • ジオAI • 金融	• 倫理 • 個人情報 • 知的財産 • 責任分解点	• 安全保障 • 主権 • リテラシー

「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・活用できる国にする」

パーパス実現のためAI基本計画等を踏まえたミッション（AISIの役割・責任）

- ① AI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を促進する（もって、国民生活の向上・社会課題の解決を促進）
- ② AI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・競争力強化を支える
- 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に貢献する

- ミッション達成に必要なAISIが行うべき4つの活動
 - （1）AIの安全性に関する物差し（ベンチマーク）を作る
 - （2）AIモデル・システムを自ら評価する能力を持つ
 - （3）AI関連情報の収集・分析・提供を率先して行う
 - （4）AI安全性に関して国際協調を主導し、国際競争力を高める

グローバルなAIセーフティの取り組みと日本

「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



① 科学的評価モデル

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



② 規制・インフラモデル

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



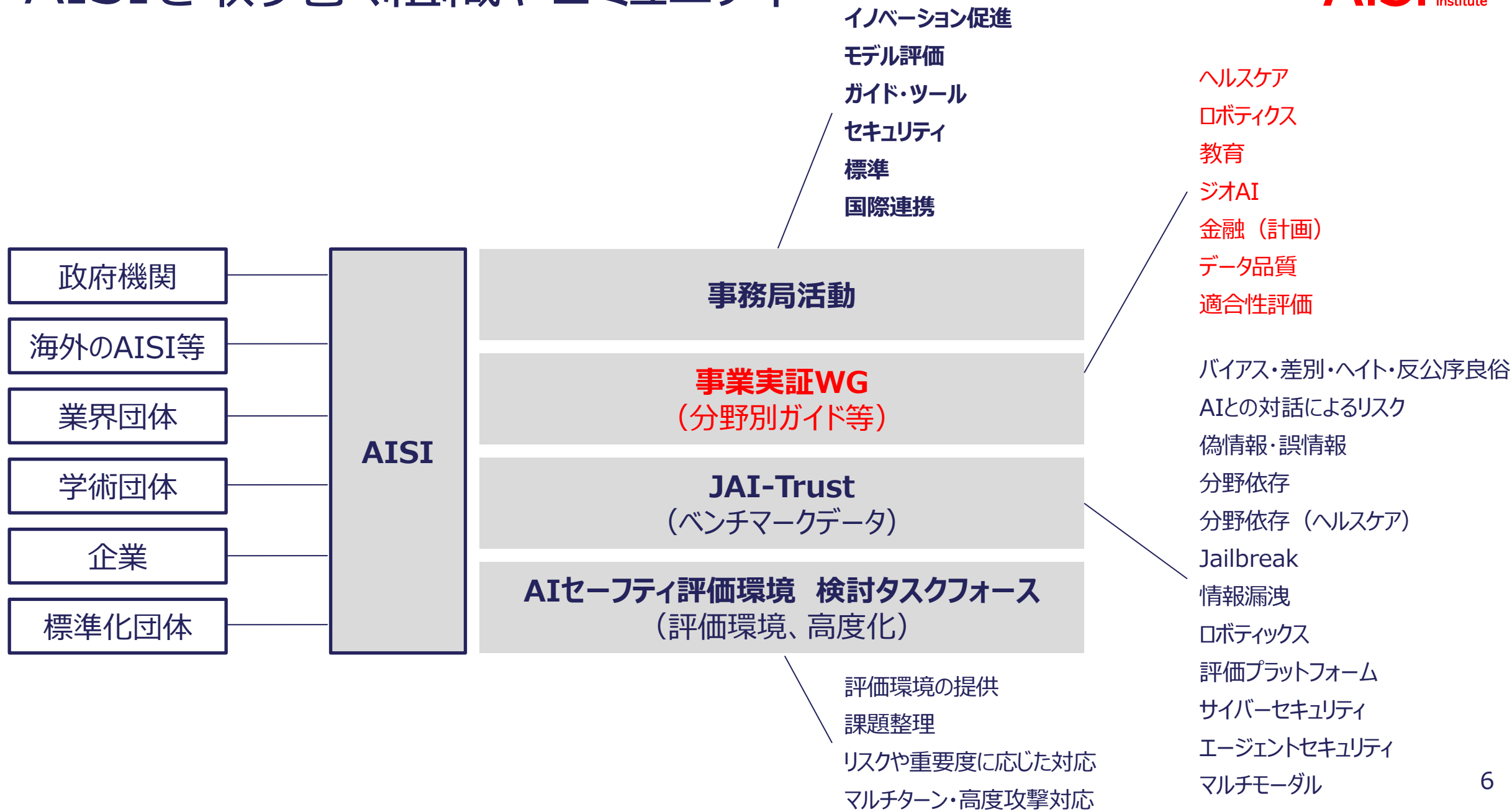
③ 実装・エコシステムモデル

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を取り巻く組織やコミュニティ



各取り組みを組み合わせAIセーフティを実現

- AIセーフティの確保には「何を見るか」を整理するだけでなく、「どのように測るか」「どのように確かめるか」が重要
- さらに、持続的な仕組みのための、現場の声（実証・運用）や人材も重要

何を見るか

AI事業者ガイドライン
AIセーフティに関する評価観点ガイド

どのように測るか
どのように確かめるか

AIモデル評価
AIセーフティ評価環境 検討タスクフォース
JAI-Trust（AIの安全性ベンチマーク構築プロジェクト）

ユースケースでの実証

事業実証ワーキンググループ

運用

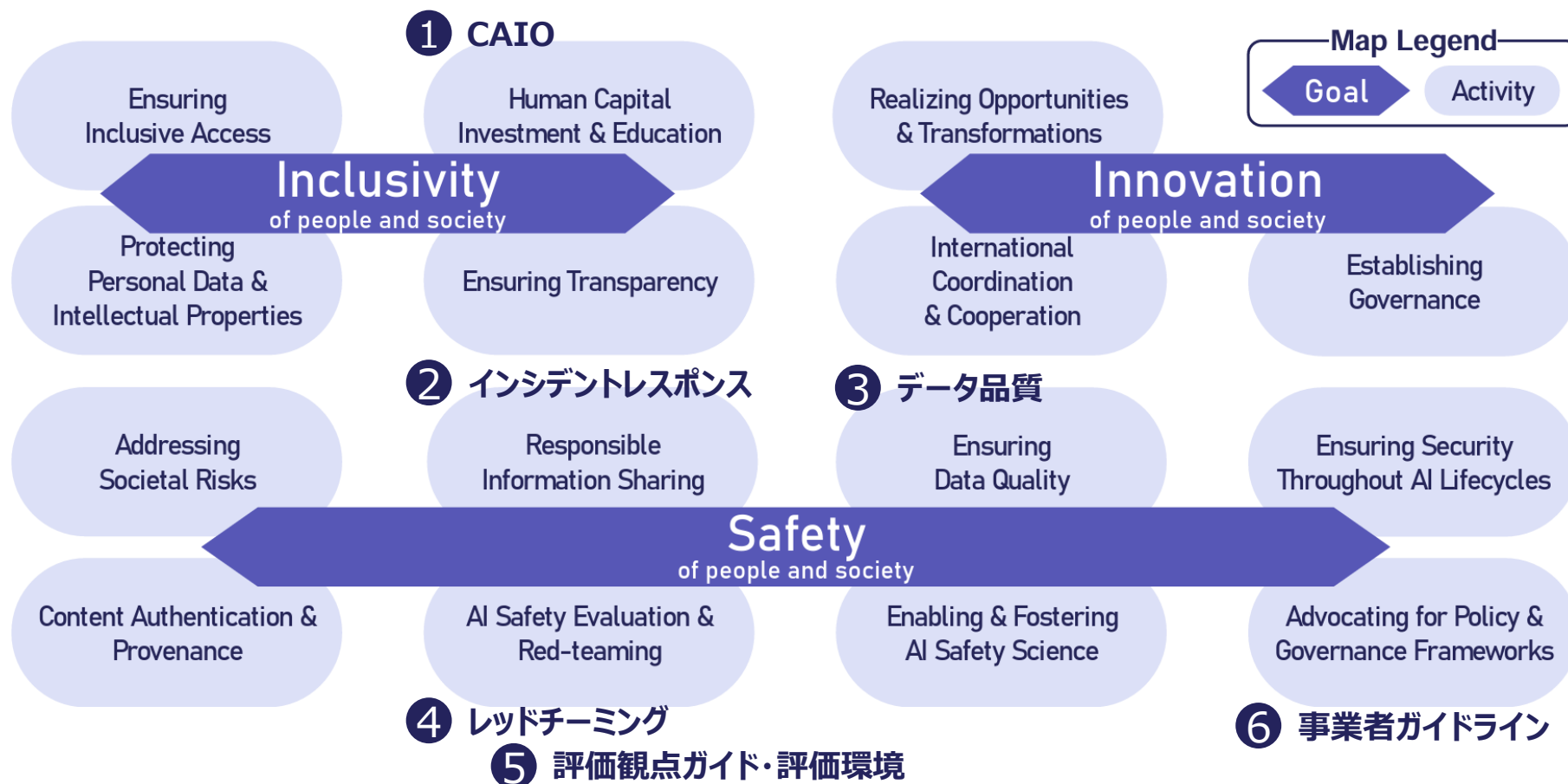
AIインシデントレスポンス（AI-IRS）

人材

CAIOガイド
各種情報・教材提供

技術マップAMAISと現在の主な取り組み

実務に適用可能な評価手法やガイドを体系的に整備することを重視



1. Chief AI Officerガイドブック

-主に民間事業者がCAIOを設置・運用する際の標準的な実務指針を提供

2. インシデントレスポンスアプローチブック

- AIシステム特有のリスクに起因するインシデントに対応するための新たな枠組みとしてアプローチブックを公開

3. データ品質マネジメントガイドブック

- データ品質を持続的に確保するため何をすべきか整理したガイド

4. レッドチーミング手法ガイド

- 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ガイド

5. 評価観点ガイド・評価環境

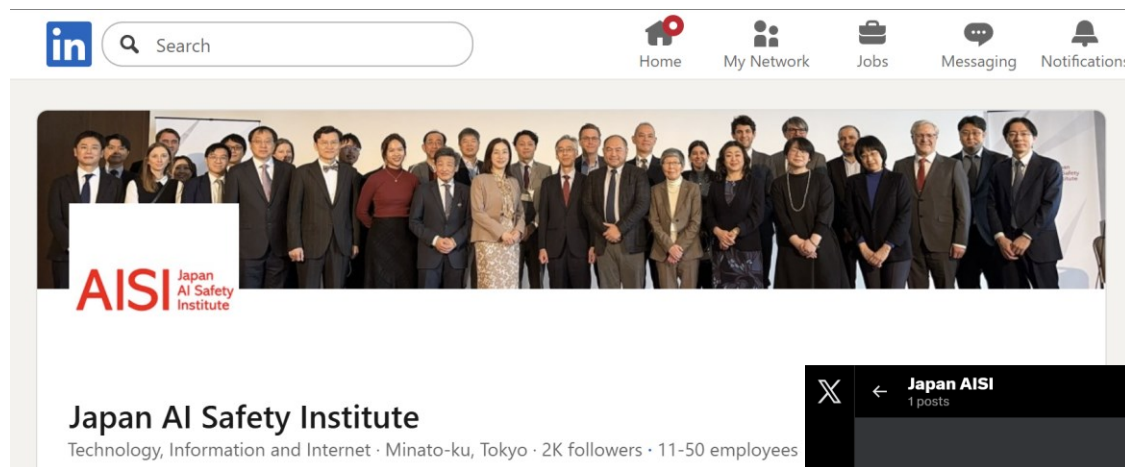
- AIシステムの安全性を評価する際の基本的な考え方を示したガイド

6. 事業者ガイドライン

-総務省・経済産業省がリビングドキュメントとして更新の検討を取りまとめ公表

情報発信の強化

Webサイト、SNS、講演などを組み合わせて、必要な人にAIセーフティの情報を届け、導入、促進を図る。



最新の活動

2026年10月6日 [AI情報通：10月5日15時の情報](#)

2026年10月2日 [AI検証ノート：高度なAIサイバー能力のオープンモデルへの拡がり](#)
[AI Experiment Notes: The Proliferation of Advanced AI Cyber Capabilities to Open Models](#)

[AI検証ノート：ExploitBenchによるClaude Opus 4.8のサイバー能力評価 ～事例分析の重要性について～](#)
[AI Experiment Notes: Cyber Capability Evaluation of Claude Opus 4.8 Using ExploitBench ~Importance of Case Analysis~](#)

AISI

Japan AI Safety Institute