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

2025-01-06

AISI Japan AI Safety Institute

Japan AI Safety Institute

AISI

AI 戦略

統合イノベーション戦略2024における３つの強化方策

AISI Japan AI Safety Institute

(1) 重要技術に関する統合的な戦略

①コア技術の開発、他の戦略分野との技術の融合による研究開発（産学官の連携、AI・ロボティクス・IoT等による研究開発推進等）

②国内産業基盤の確立、スタートアップ等によるイノベーション促進（ユースケースの早期創出、拠点・ハブ機能の強化等）

③産学官を挙げた人材の育成・確保（産業化を担う人材、市場開拓を担う人材、研究開発を担う人材の育成・確保等）

(2) グローバルな視点での連携強化

①重要技術等に関する国際的なルールメイキングの主導・参画（開発・利用の促進、安全性確保、プレゼンスの確保等）

②科学技術・イノベーション政策と経済安全保障政策との連携強化（国際協力・国際連携を含めた戦略的な研究開発、技術流出防止等）

③グローバルな視点でのリソースの積極活用、戦略的な協働（国際頭脳循環の拠点形成、国際科学トップサークルへの参画等）

(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、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）

(3) AI分野の競争力強化と安全・安心の確保

AISI Japan AI Safety Institute

- 生成AIはインターネットにも匹敵する技術革新とされ、社会経済システムに大きな変革をもたらす一方で、偽・誤情報の流布や犯罪の巧妙化など様々なリスクも指摘され、安全・安心の確保が求められる。
- 米国企業等の高性能・大規模な汎用基盤モデルが先行する中、我が国もそれに追随すべく計算資源の整備や大規模モデルの開発が進んでおり、また、小規模・高性能なモデルや複数モデルの組合せの開発など、新たな研究も進んでいる。
- AIはあらゆる分野で利用され、AIの開発や利活用等のイノベーションが社会課題の解決や我が国の競争力に直結する可能性がある。我が国においては、生成AIを含むAIの様々なリスクを抑え、安全・安心な環境を確保しつつ、イノベーションを加速する好循環の形成を図っていく。加えて、我が国が主導する広島AIプロセス等を通じて、今後も国際的にリーダーシップを発揮していく。

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（データ整備含む）
- AI利活用の推進
- インフラの高度化
- 人材の育成・確保

② AIの安全・安心の確保
- 自発的ガバナンスと制度の検討
- AIの安全性の検討
- 偽・誤情報への対策
- 知的財産権等

③ 国際的な連携・協調の推進

統合イノベーション戦略2024.pdf

Japan AI Safety Institute

AISI

AISI(AIセーフティ・インスティテュート)

日本におけるAISIの設立

AISI Japan AI Safety Institute

- 2023年5月
  - 岸田総理大臣(当時)が「広島AIプロセス（※1）」を提唱
    - G7広島サミットで提唱された生成AIに関する国際的なルールの検討を行うためのプロセス
- 2023年10月
  - 広島AIプロセス「国際指針」及び「国際行動規範」に合意
    - 生成AIを含む高度なAIシステムに関する国際的な指針と行動規範
- 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14日**
  - IPA（情報処理推進機構）にAIセーフティ・インスティテュート（AISI）を設立

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

各国のAI安全性確保への取組み

AISI Japan AI Safety Institute

- **米国**
  - NIST（国立標準技術研究所）にAISIを設立
  - 基本は民間主導、民間企業とのコンソーシアム（AISIC）との協働を強力に推進
  - 人員規模は30人程度。80名位を目指し推進中
- **英国**
  - DSIT（科学イノベーション技術省）にAISIを設立
  - 政府主導で、AIの安全性に関する評価やTestingを強力に推進
  - 規模は100名体制。技術者を多数雇用予定。また、サンフランシスコオフィスを開業
- **EU**
  - EC（欧州委員会）にあるAIオフィスで、利活用に加え、安全性も推進。AI法の整備と推進も担う
  - 60人程度の規模
- **シンガポール**
  - 南洋理工大学（NTU）内のデジタルトラストセンターがシンガポールのAISIを指定
  - 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
- **カナダ**
  - 国内機関の協力のもとAISI設立
- **韓国**
  - 2024年11月 AISI設立
  - アジアのハブを目指す
- **オーストラリア**
  - 国立の研究所がAISI機能を担う

AISIの概要

AISI Japan AI Safety Institute

- AISIの位置づけ
  - 今後、官民が協力して、AIの安全安心な活用が促進されるよう、AIの開発や利用をする全ての関係者がAIのリスクを正しく認識し、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ようにしていく必要がある。
  - また、これらの取組を通じ、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していく必要がある。
  - AISIは、上記を実現するための**官民の取組を支援する機関**である。

- 取組方針
  - 技術がグローバルかつ目まぐるしく進歩していることから、国内、国際的な関係機関と協調して取組を推進していく。

AISIの役割とスコープ

AISI Japan AI Safety Institute

- 役割
  - 政府への支援として、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の支援を行うとともに、日本におけるAIセーフティのハブとして、産学における関連取組の最新情報を集約し、関係企業・団体間の連携を促進し、さらに、 他国のAIセーフティ関係機関との連携により国際的なコンセンサスを構築する。
    - 自ら研究開発する組織ではない

- 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
    - AIシステム
    - ガバナンス
    - コンテンツ
    - データ

実現に向けた業務

AISI Japan AI Safety Institute

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AI Safety Institute等)との国際連携に関する業務

直近の活動と成果物

AISI Japan AI Safety Institute

	国際	AISI		政府
2024	イベント	成果物	効果	
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**		✓ セーフティに関する基礎情報を提供 ✓ グローバルな確認が可能になる ✓ 評価のポイントがわかる ✓ テスト手法がわかる	• **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)		
9月		• **日米クロスウォーク2**の成果公表(9/18)		
10月		• **レッドチーミング手法ガイド**※の公表(9/25)		
11月	**AISI国際ネットワーク会合, 米国**			
12月				

今後の予定（2025年）

1月：年次レポートの公表
2月：AI アクション サミット, フランス
3月：評価観点ガイド・レッドチーミング手法ガイドの更新、調査事業のとりまとめ、AI事業者ガイドラインの更新

※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

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］

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）

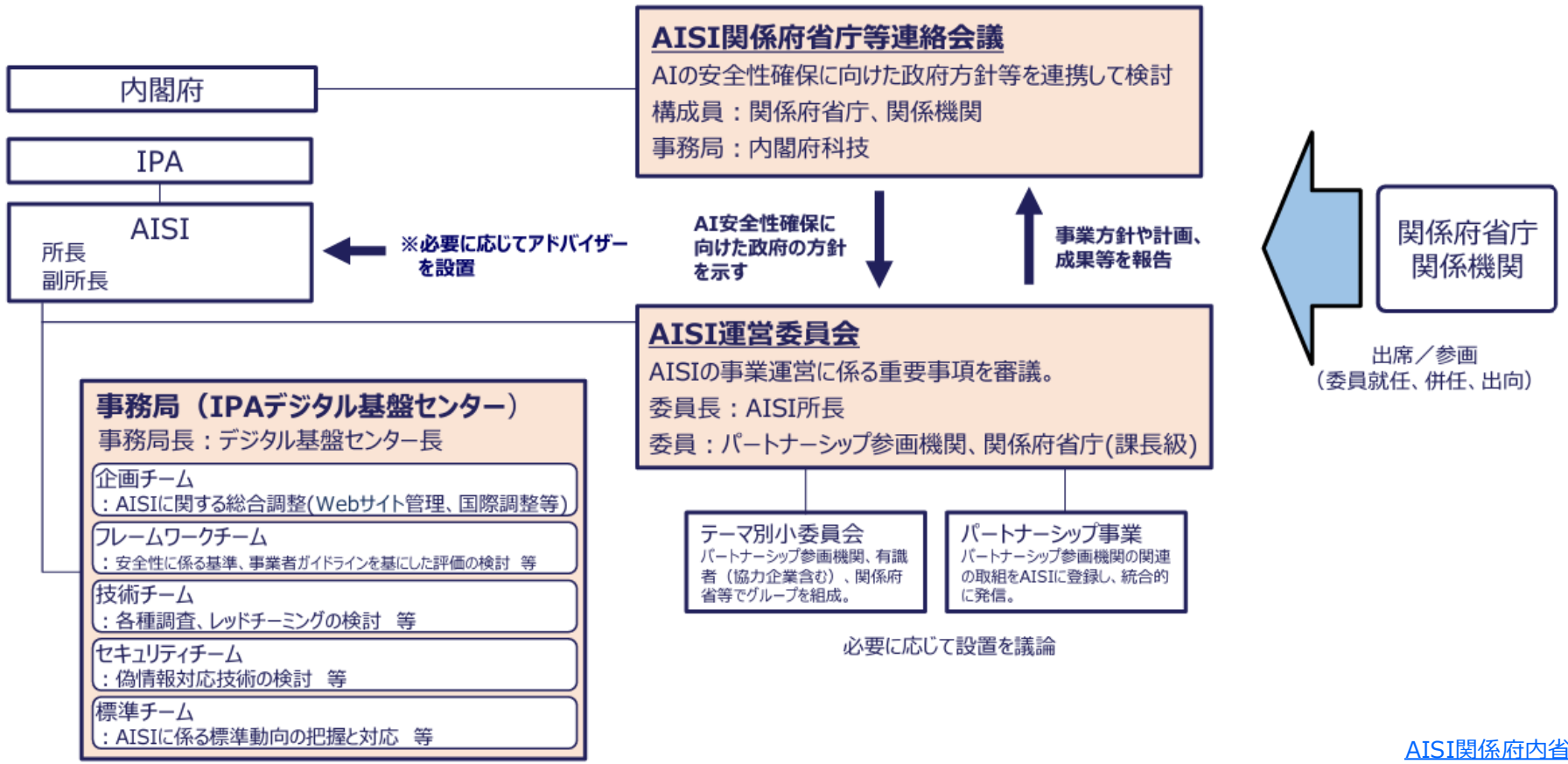
副所長

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」を設置し、重要事項を審議（年間２～３回の開催を予定）。AISIの中に、AISI所長を委員長とする「AISI運営委員会」を設置（月１回の開催を予定）。
  - 運営委員会の下に、必要に応じて、「テーマ別小委員会」や「パートナーシップ事業」（研究機関等の関連の取組みをAISI事業として発信）を設置。



【関係府省庁】
・内閣府（科学技術・イノベーション推進事務局）
・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
・経産省
・防衛省

【関係機関】
・情報処理推進機構（IPA）
・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

AISI関係府内省庁等連絡会議(第1回).pdf

AISIパートナーシップ協定の発効

AISI Japan AI Safety Institute

- 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。このため、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）を2024年8月より発効。

  - 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

日本AIセーフティ・インスティチュートパートナーシップ協定.pdf

AISIでの実施予定事項（１）

AISI Japan AI Safety Institute

- 企画
  - AISIの戦略や計画を作成、予算を管理
  - AIセーフティに関する状況把握
  - 広報（AISIサイト管理含む）
  - 採用、人材育成（教材作成含む）
  - 関係機関（国際含む）との調整・支援
- フレームワーク
  - AI事業者ガイドラインの作成等（総務省・経産省）の支援（例：クロスウォーク）
  - AIリスク管理のフレームワークの国際調整支援
  - AIガバナンスに関わる国内外の資料の収集とその結果に基づく技術的助言
  - 認証・認定の在り方の検討支援

AISIでの実施予定事項（２）

AISI Japan AI Safety Institute

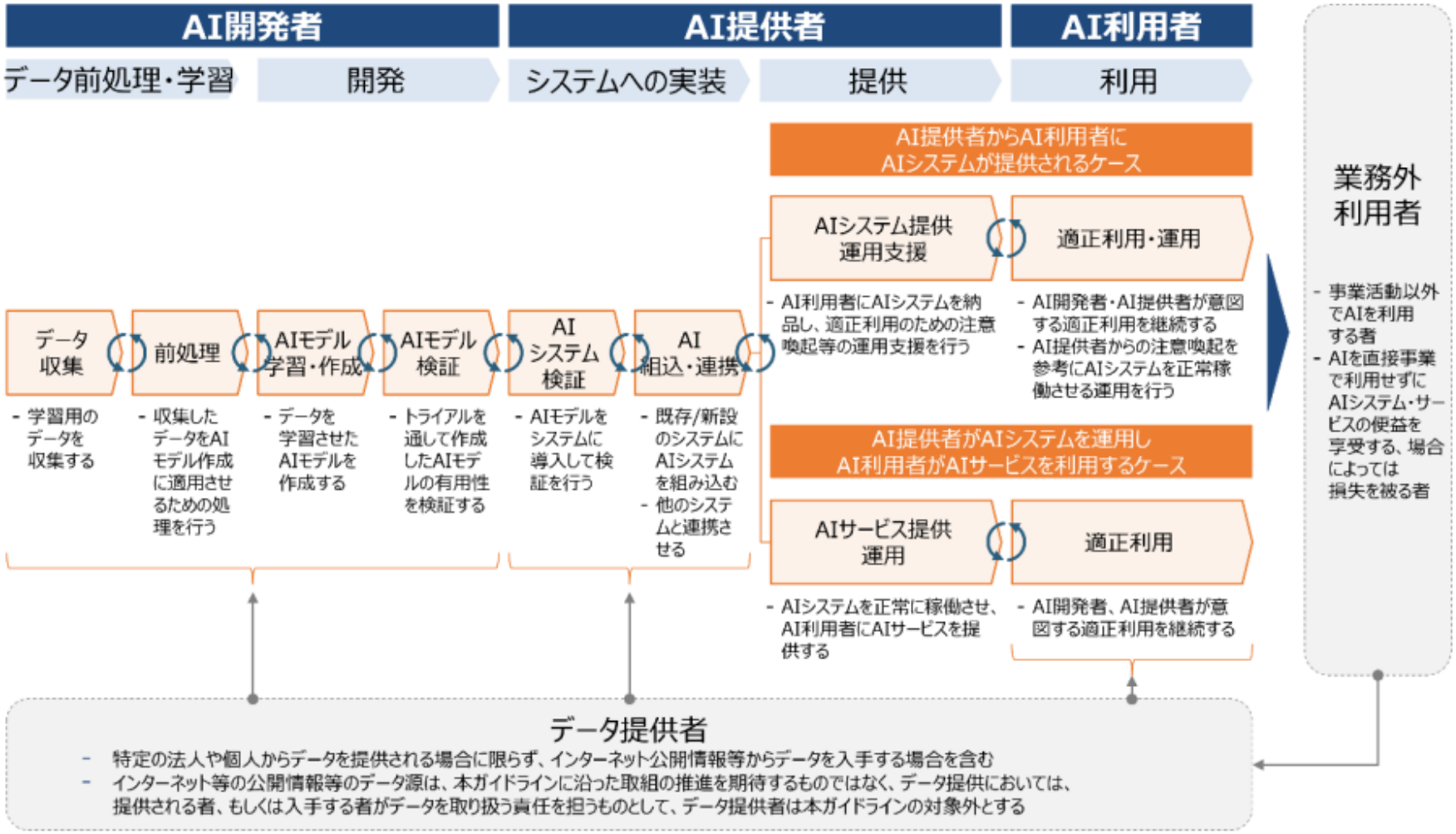
- 技術
  - 技術企画
  - レッドチームの実施方法の整理
  - 技術関連の基準、ガイドラインの整備等に資する情報収集や助言
    - 合成コンテンツ・偽情報・誤情報
    - バイアス、データチェック、来歴管理
  - テストベッド等の必要ツールの検討
- セキュリティ
  - AIに対するセキュリティ対策の検討
  - AIを使ったセキュリティ事象への対策支援
- 標準
  - ISO　SC42の推進（産総研）の支援
  - その他標準情報の収集

AISI関連活動の成果実現に向けた直近の取組

AI事業者ガイドラインの概要

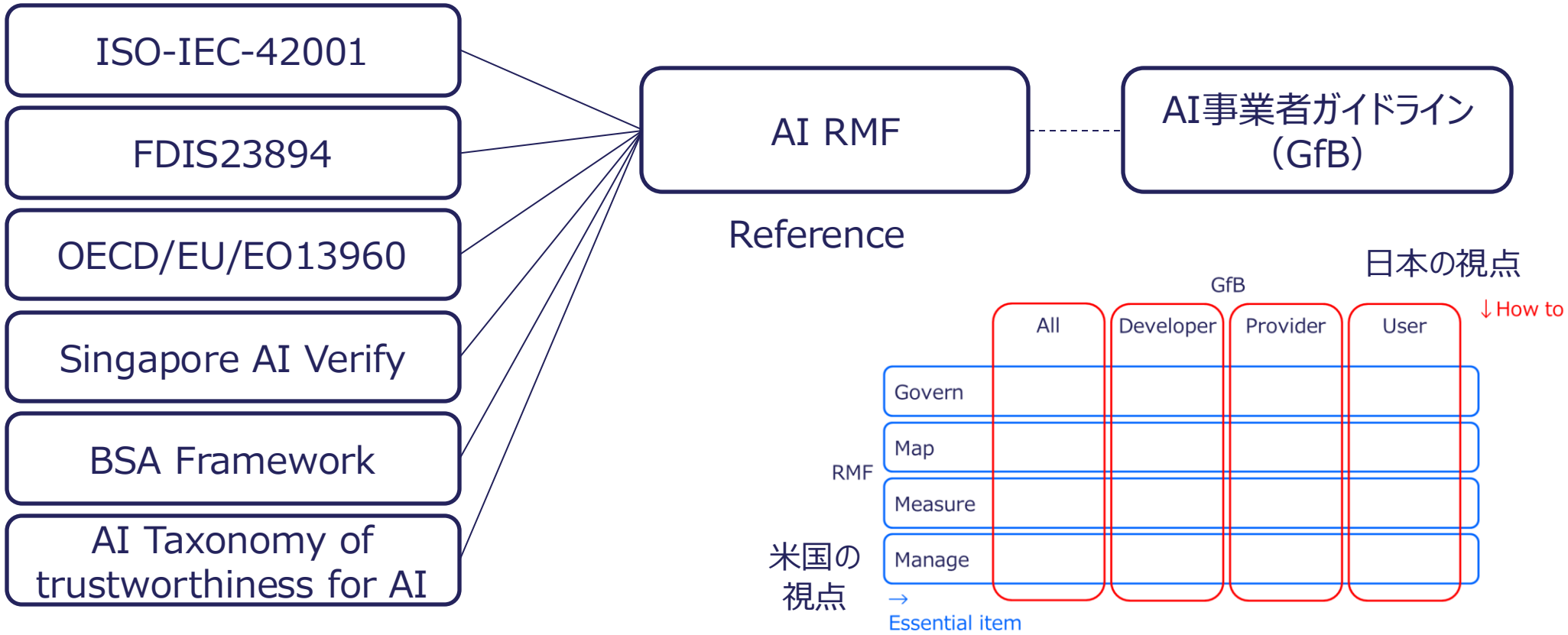
AISI Japan AI Safety Institute

- 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



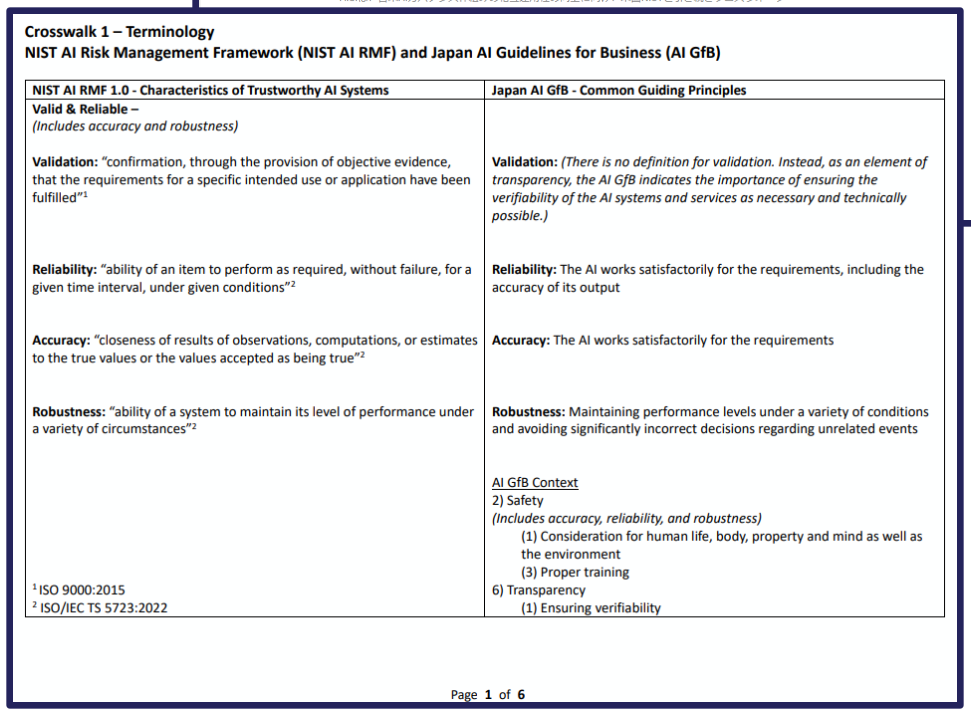
日米クロスウォークの概要

- 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



日米クロスウォークの成果

- クロスウォーク１の成果を公開（4月30日）
  Crosswalk1.pdf
- クロスウォーク２の成果を公開（9月18日）
  Crosswalk2.pdf



Crosswalk 1 – Terminology
NIST AI Risk Management Framework (NIST AI RMF) and Japan AI Guidelines for Business (AI GfB)

NIST AI RMF 1.0 - Characteristics of Trustworthy AI Systems	Japan AI GfB - Common Guiding Principles
Valid & Reliable – *(Includes accuracy and robustness)*	
Validation: "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled"[1]	**Validation:** *(There is no definition for validation. Instead, as an element of transparency, the AI GfB indicates the importance of ensuring the verifiability of the AI systems and services as necessary and technically possible.)*
Reliability: "ability of an item to perform as required, without failure, for a given time interval, under given conditions"[2]	**Reliability:** The AI works satisfactorily for the requirements, including the accuracy of its output
Accuracy: "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true"[2]	**Accuracy:** The AI works satisfactorily for the requirements
Robustness: "ability of a system to maintain its level of performance under a variety of circumstances"[2]	**Robustness:** Maintaining performance levels under a variety of conditions and avoiding significantly incorrect decisions regarding unrelated events
	AI GfB Context 2) Safety *(Includes accuracy, reliability, and robustness)* (1) Consideration for human life, body, property and mind as well as the environment (3) Proper training 6) Transparency (1) Ensuring verifiability

[1] ISO 9000:2015
[2] ISO/IEC TS 5723:2022

Page 1 of 6

評価観点ガイドの公開

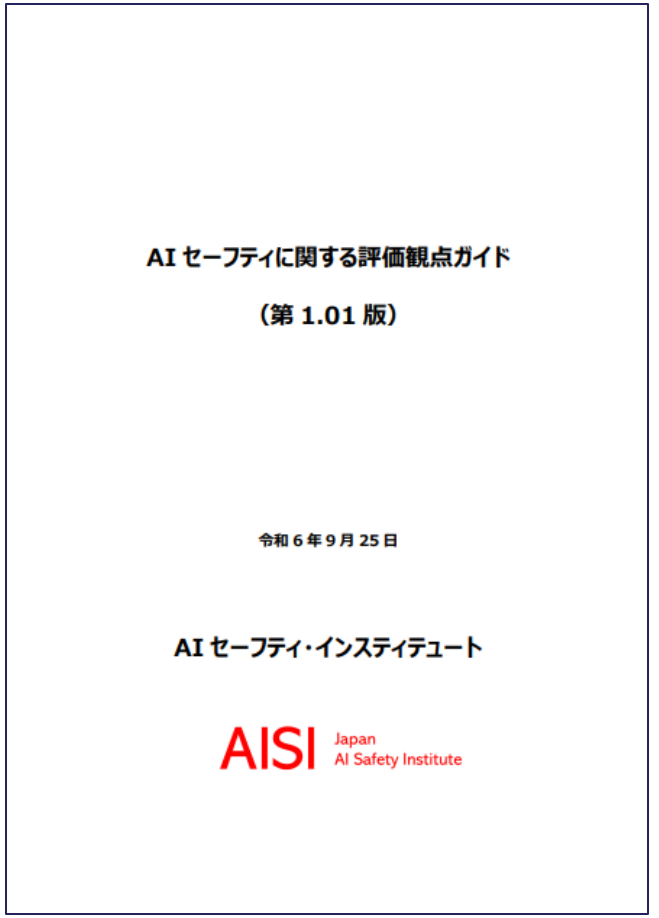
AIセーフティに関する評価観点ガイドの意義と活用法について

- AIセーフティに関する評価観点ガイドは、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であり、事業者がAIを開発・提供する際の参考とするもの。

- 具体的には、
  - 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

  などが記載されている。

- 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。

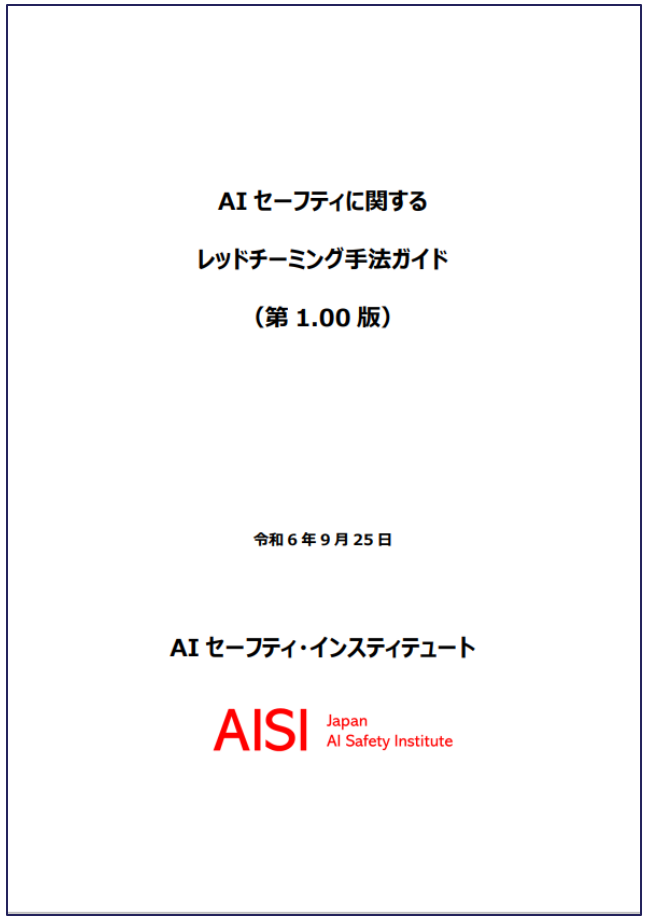


AIセーフティに関する評価観点ガイド概要.pdf

レッドチーミング手法ガイドの公開

AIセーフティに関するレッドチーミング手法ガイドの意義と活用法について

- このガイドは、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であり、事業者がAIを開発・提供する際の参考とするもの。
- 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.pdf

国際連携

AISI Japan AI Safety Institute

- **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（スタンフォード、4月16日）
    - 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（ソウル、5月21-22日）
    - 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
  - シンガポールのアジアTech xサミット（オンライン、5月31日）
    - 米国AISIの所長等とパネルディスカッション
  - 国連未来サミット（国連本部、9月22日）
  - 国連Global Compact Leaders Summit 2024 （国連本部、9月24日）
    - 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（サンフランシスコ、11月10-11日）
- **各国との意見交換**

  AI関連事業者及び団体との事務レベルの意見交換を積極的に実施
  - 米国、英国、EU、シンガポール、オーストラリア、韓国との意見交換
  - 事業者等のエグゼクティブとの意見交換
  - GPAIワークショップ（パリ）参加（事務局、5月22・23日）

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

今後の取り組み予定

AISI Japan AI Safety Institute

- 民間企業との協力関係
  - 「事業実証ワーキンググループ（ＷＧ）」の設置を検討
    - 第３回AISI運営委員会にて公表
- 調査研究の拡大
  - 評価観点ガイドやレッドチーミング手法ガイドの改定に向けた調査
    - 対象をマルチモーダル基盤モデルに拡大
  - 事業実証ＷＧの取組に向けた調査
    - AIセーフティの評価環境の一部を先行して構築し、WG活動の環境を整備
  - AIセーフティの自動評価に関する調査
    - AIセーフティ評価を広く一般化するため、評価の自動化/省力化を検討
- 直近の主要イベント
  - 2025年　2月10-11日(パリ)　　Artificial Intelligence Action Summit

AISI

Japan AI Safety Institute