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

2024-11-01

AISI Japan AI Safety Institute

日本におけるAISIの設立

AISI Japan AI Safety Institute

- 2023年5月
  - 岸田総理大臣が「広島AIプロセス（※）」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

所長　村上　明子

出典：
広島AIプロセス <https://www.soumu.go.jp/hiroshimaaiprocess/documents.html>
AI Safety Summit 2023 <https://www.gov.uk/government/topical-events/ai-safety-summit-2023>
AI戦略会議 <https://www.kantei.go.jp/jp/101_kishida/actions/202312/21ai.html>
AIセーフティ・インスティテュート <https://aisi.go.jp/>

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、今夏にサンフランシスコオフィスを開業予定
- **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
- **カナダ**
  - 国内機関の協力のもとAISI設立
- **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
- **オーストラリア**
  - 国立の研究所がAISI機能を担う
- **韓国**
  - 2024年度末を目指しAISIを準備中
  - アジアのハブを目指す

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連携する。
    - 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
    - ガバナンス
    - AIシステム
    - コンテンツ
    - データ

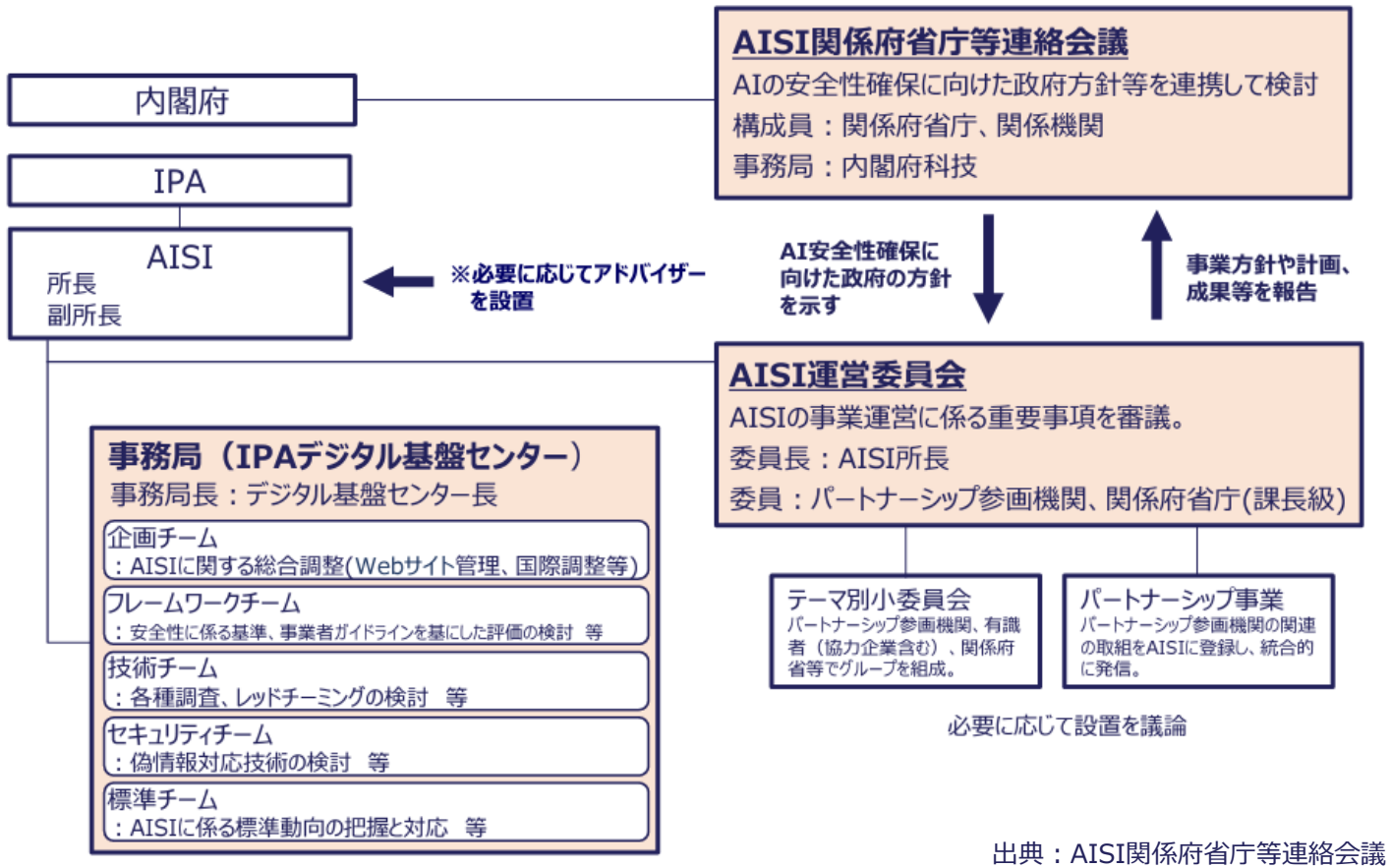
実現に向けた業務

AISI Japan AI Safety Institute

1. 安全性評価に係る調査、基準等の検討
   ① 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
   ② 安全性に係る基準、ガイダンス等の検討
   ③ 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

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

出典：AISI関係府省庁等連絡会議（第1回/令和6年2月29日）資料２

関係機関とのパートナーシップ

AISI Japan AI Safety Institute

1. AIの安全性評価に関する取組を進めていく上では、IPA内に設置したAISIのみならず、関係府省庁や研究開発等の関係機関のご協力を頂くことが不可欠です。

2. また、今後、各国のAISI等の機関と連携、調整を行っていくにあたっても、国内の関係府省庁、関係機関のご協力を得て、進めていくことが必要と考えています。

3. このため、関係府省庁、関係機関が連携してAIの安全性評価に係る取り組みを推進していくため、AISIからの呼びかけで、関係機関との間でパートナーシップ協定を締結していきたいと考えています。

4. 関係機関については、当面は、AISI関係府省庁等連絡会議のメンバーである、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を想定しています。

5. パートナーシップ事業に基づき行う事業については、AISIの名称で発信していくとともに、パートナーシップ参加機関もAISIの名称を使用し、ダブルクレジットで情報を発信。

当面の活動と成果予定物

AISI Japan AI Safety Institute

2024	国際	AISI		政府
	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	
7月		• 米国AI RMF 日本語翻訳版の公開(7/4)		• 各種の政府戦略の公表
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)**		
9月		• **レッドチーミング※手法ガイドの公表(9/25)**		
10月	AISI国際ネットワーク会合, 米国			
⋮				

※セキュリティの専門家が攻撃チームを作り、顧客企業に対して物理（Physical）/人(Human)/サイバー(Cyber)を組み合わせ、物理／仮想を問わず、現実に近い各種攻撃を仕掛け、企業のセキュリティ対策の実効性を検証すること

AISIでの実施予定事項１

- 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

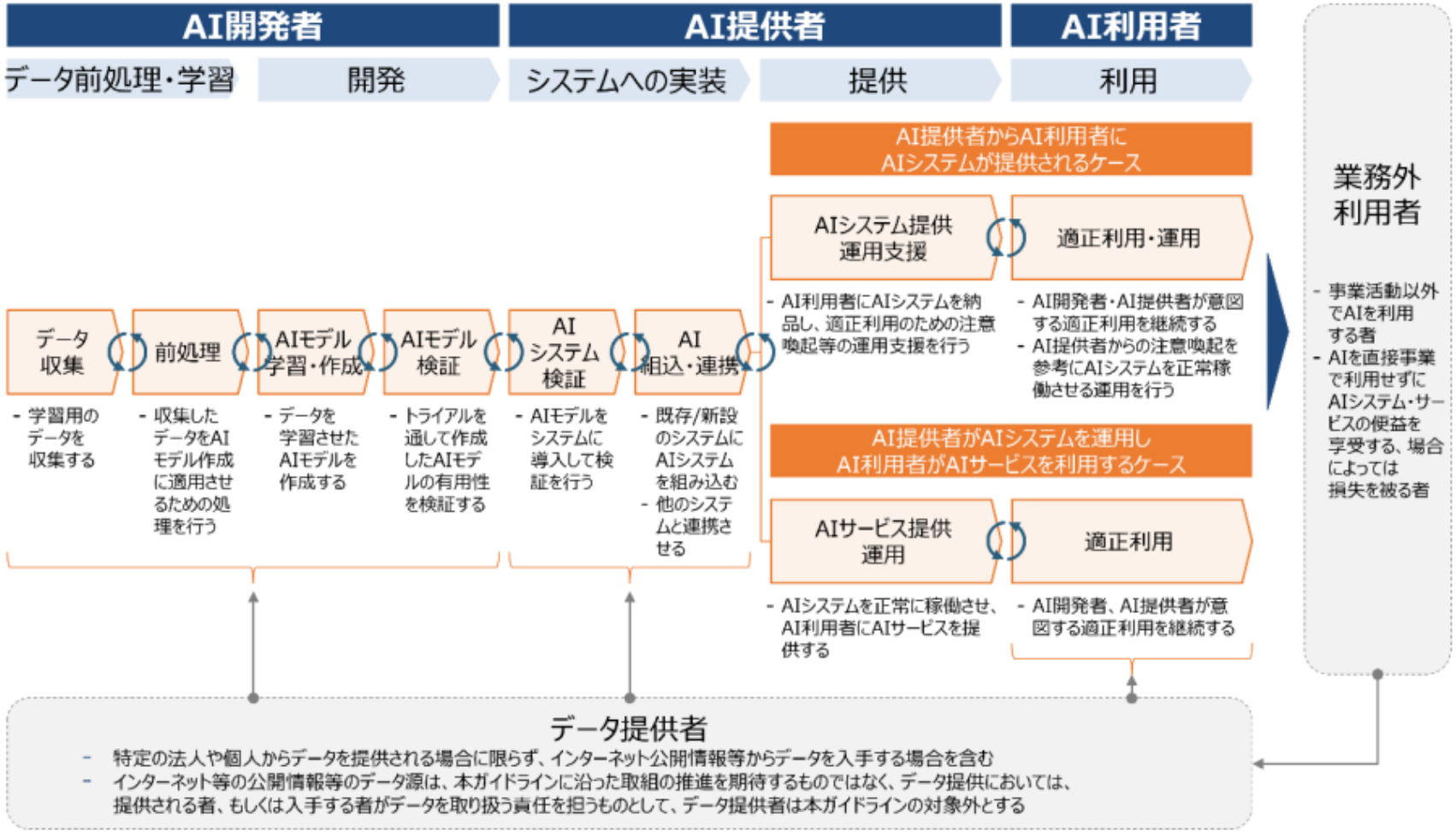
AISIでの実施予定事項２

- 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

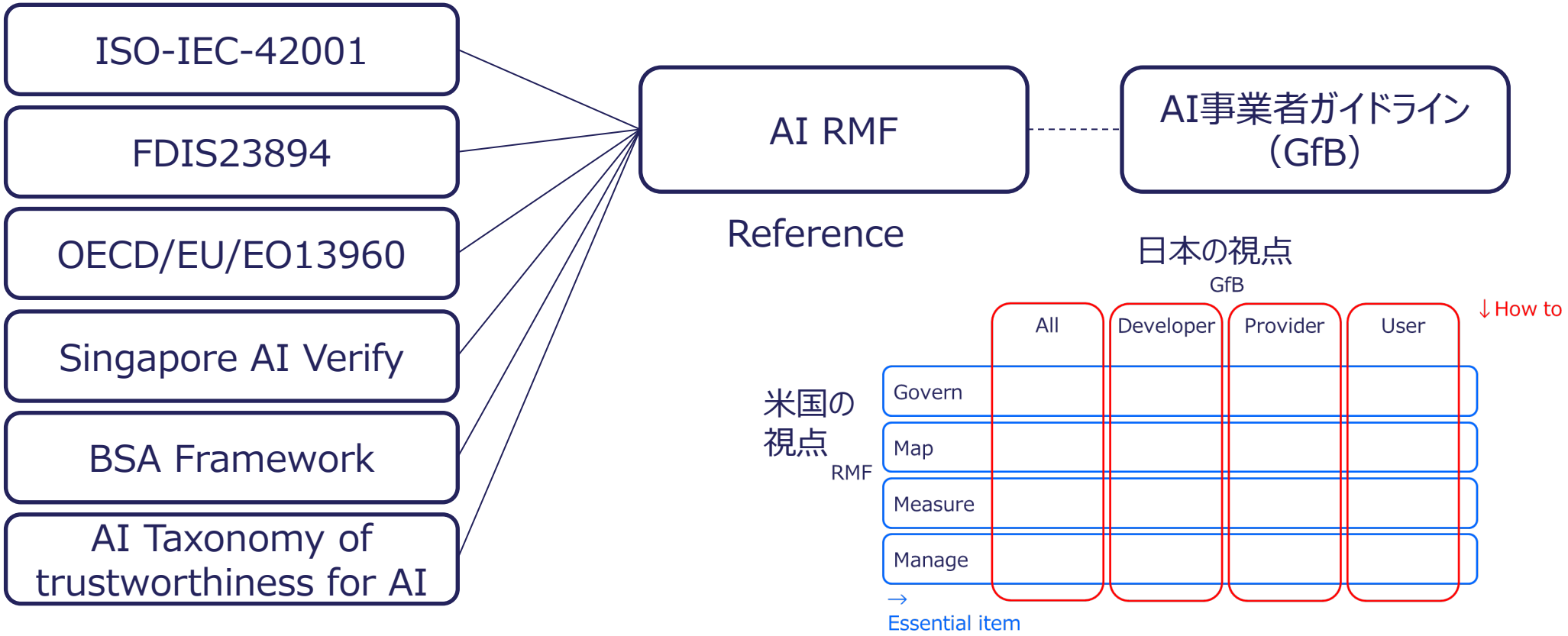
AI事業者ガイドラインの概要

- 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



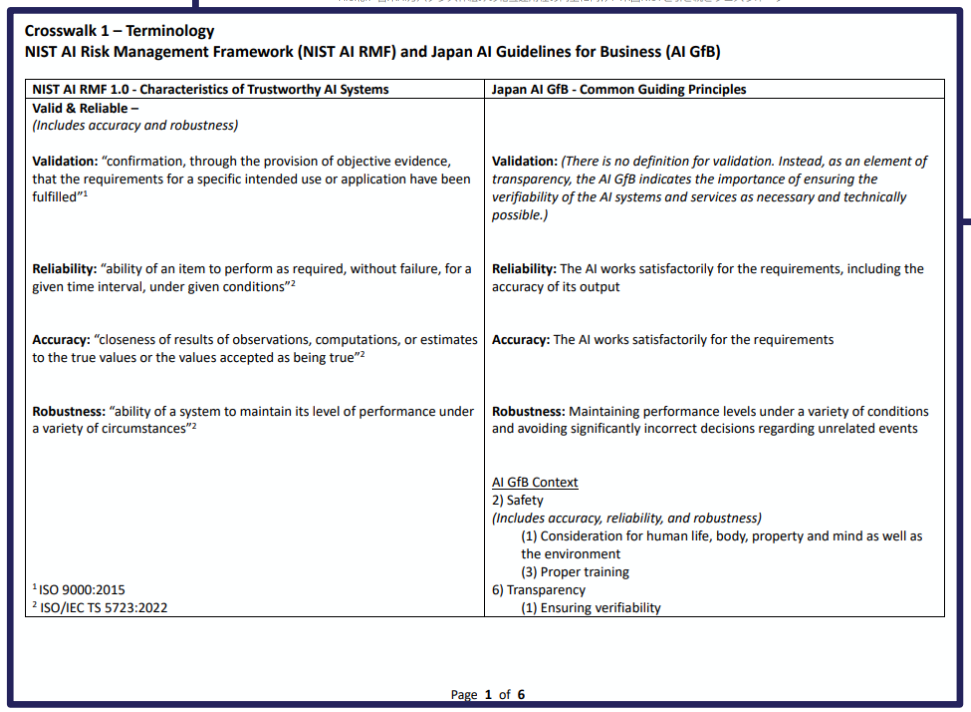
日米クロスウォークの概要

- 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）
  - 経産省、米国NISTでもツイート
- クロスウォーク２の成果を公開（9月18日）



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

AISI Japan AI Safety Institute

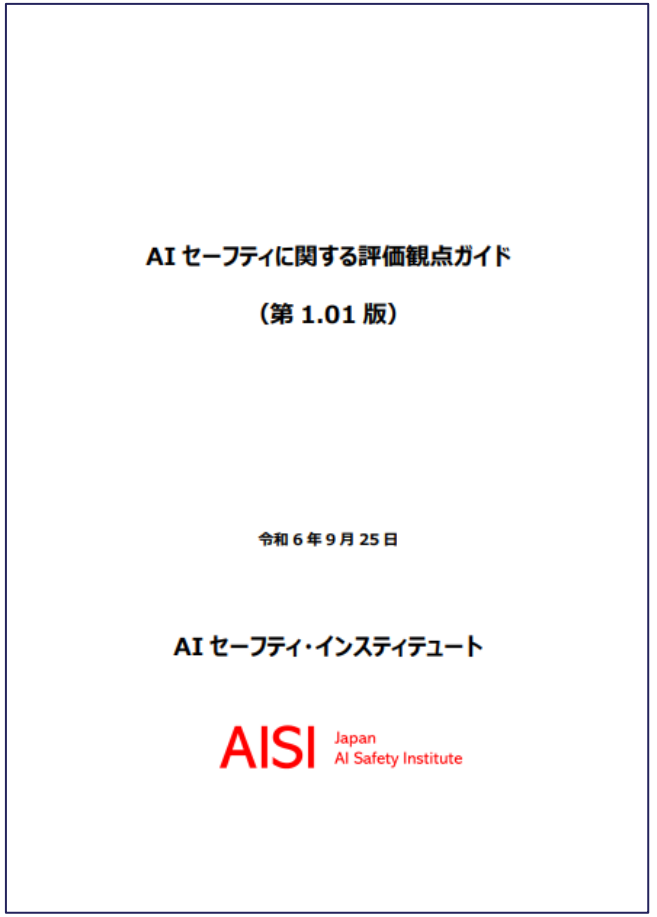
■ ＡＩセーフティに関する評価観点ガイドの意義と活用法について。

➢ 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

（第 1.01 版）

令和 6 年 9 月 25 日

AI セーフティ・インスティテュート

AISI Japan AI Safety Institute

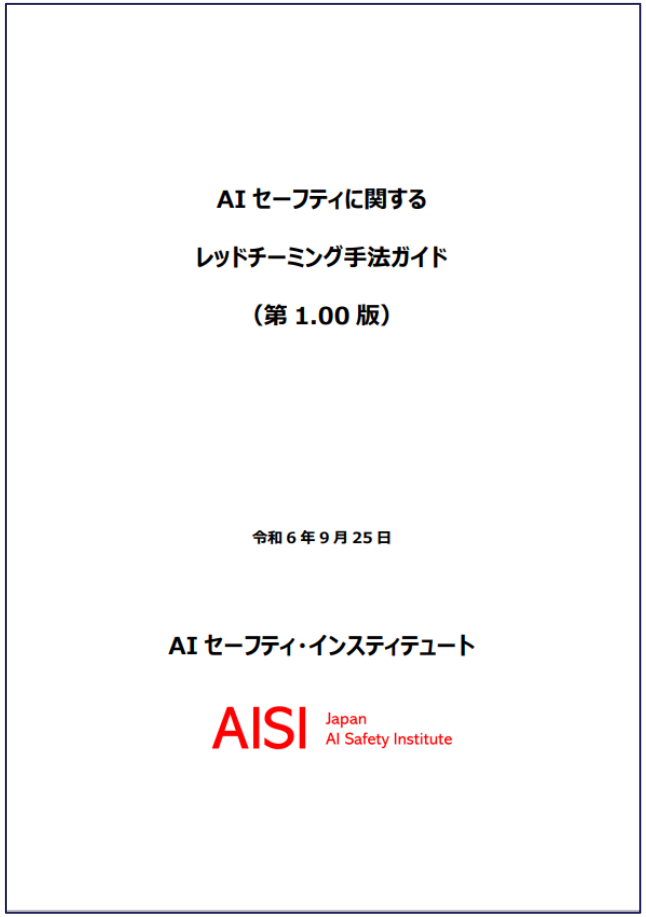
https://aisi.go.jp/assets/pdf/ai_safety_eval_summary_v1.00_ja.pdf

レッドチーミング手法ガイドの公開

AISI Japan AI Safety Institute

- ■ 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。

https://aisi.go.jp/assets/pdf/ai_safety_RT_summary_v1.00_ja.pdf



国際連携

- **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、9月22日）

    国連Global Compact Leaders Summit 2024 （国連、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
- **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

- 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
- AISIが参加する主要イベント（直近）
  - 11月8日　ISSスクエア水平ワークショップ
  - 11月11日　多文化・多言語対応の安全な大規模言語モデルの構築を目指して
  - 11月20-21日　International Network of AISIs Convening

AISI

Japan AI Safety Institute