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

AISI Japan AI Safety Institute

AISIの概要

- 2024年2月14日設立

- 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

- 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

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

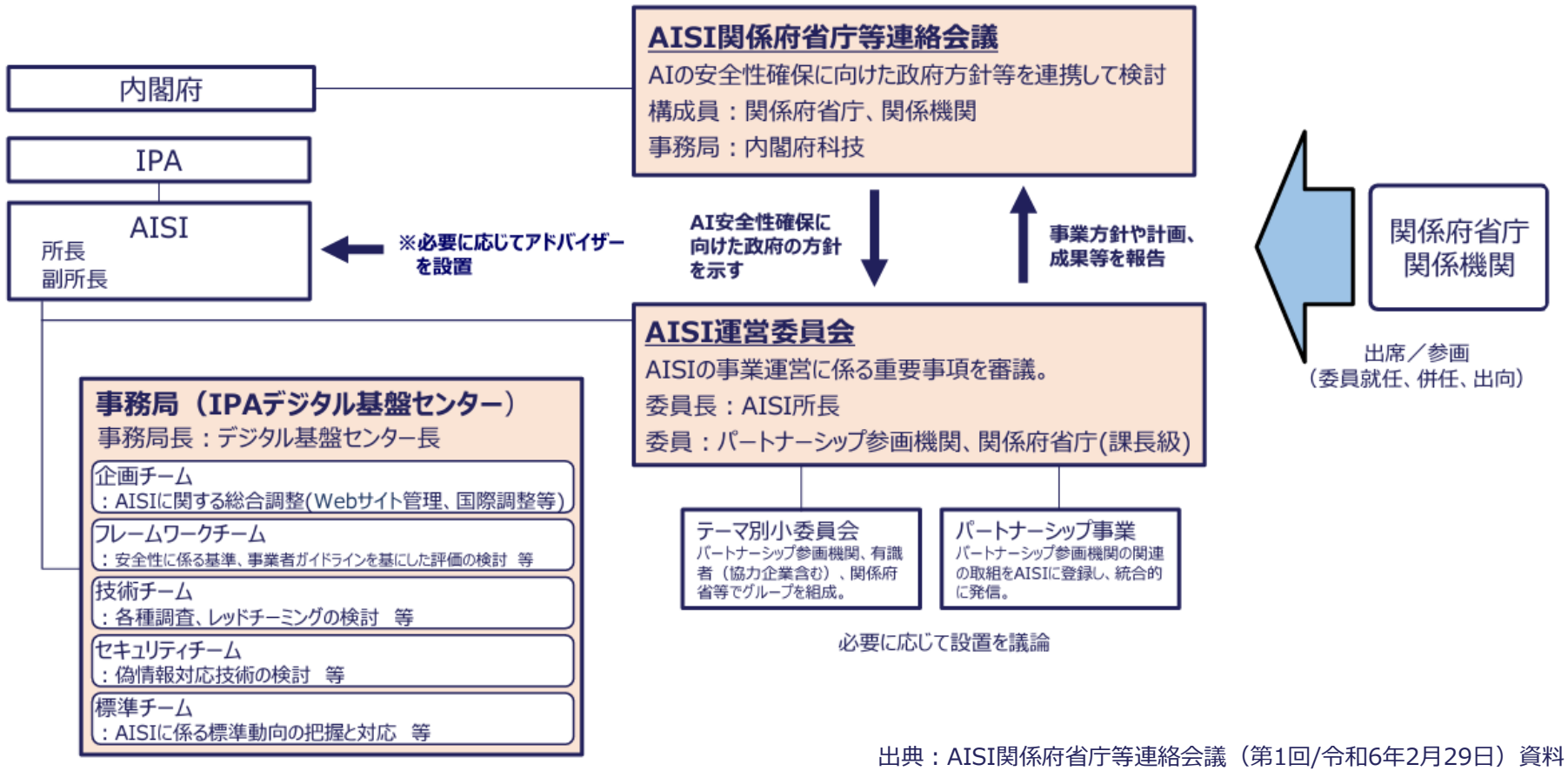
実現に向けた業務

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。

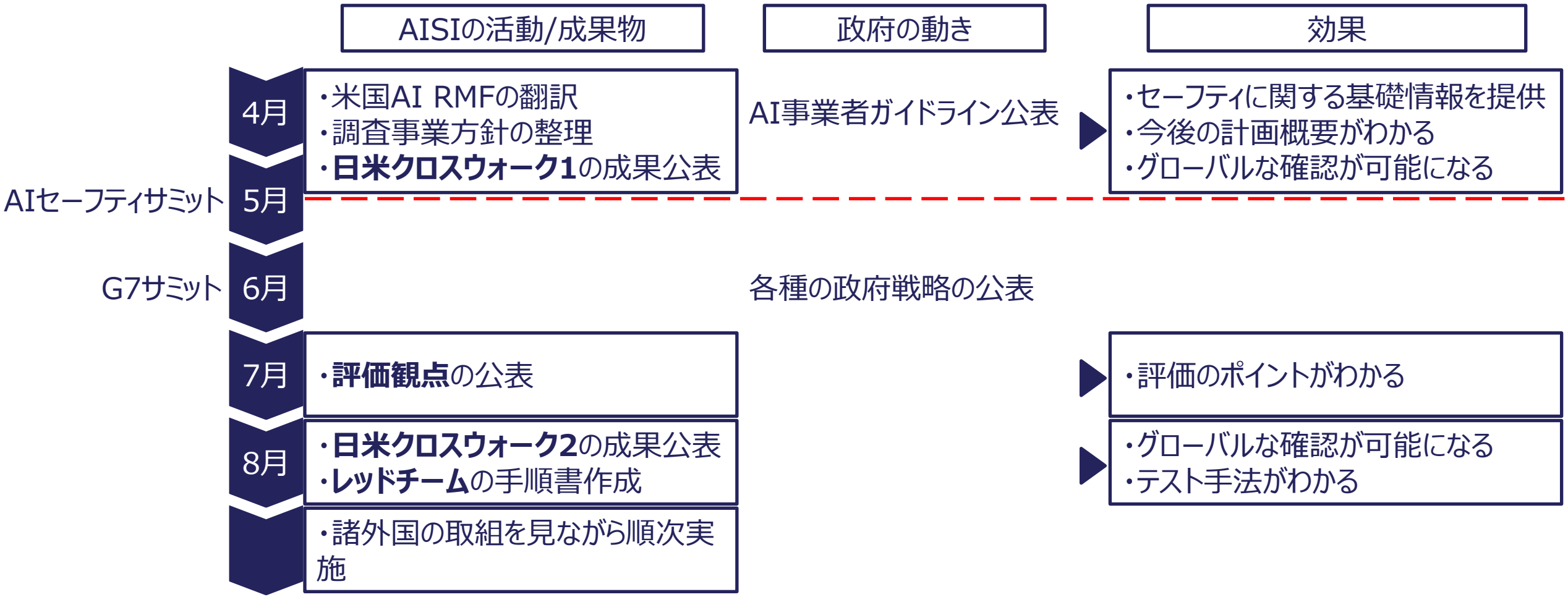


【関係府省庁】
・内閣府（科学技術・イノベーション推進事務局）
・国家安全保障局
・内閣サイバーセキュリティセンター
・警察庁
・デジタル庁
・総務省
・外務省
・文科省
・経産省
・防衛省

【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

当面の活動と成果予定物



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

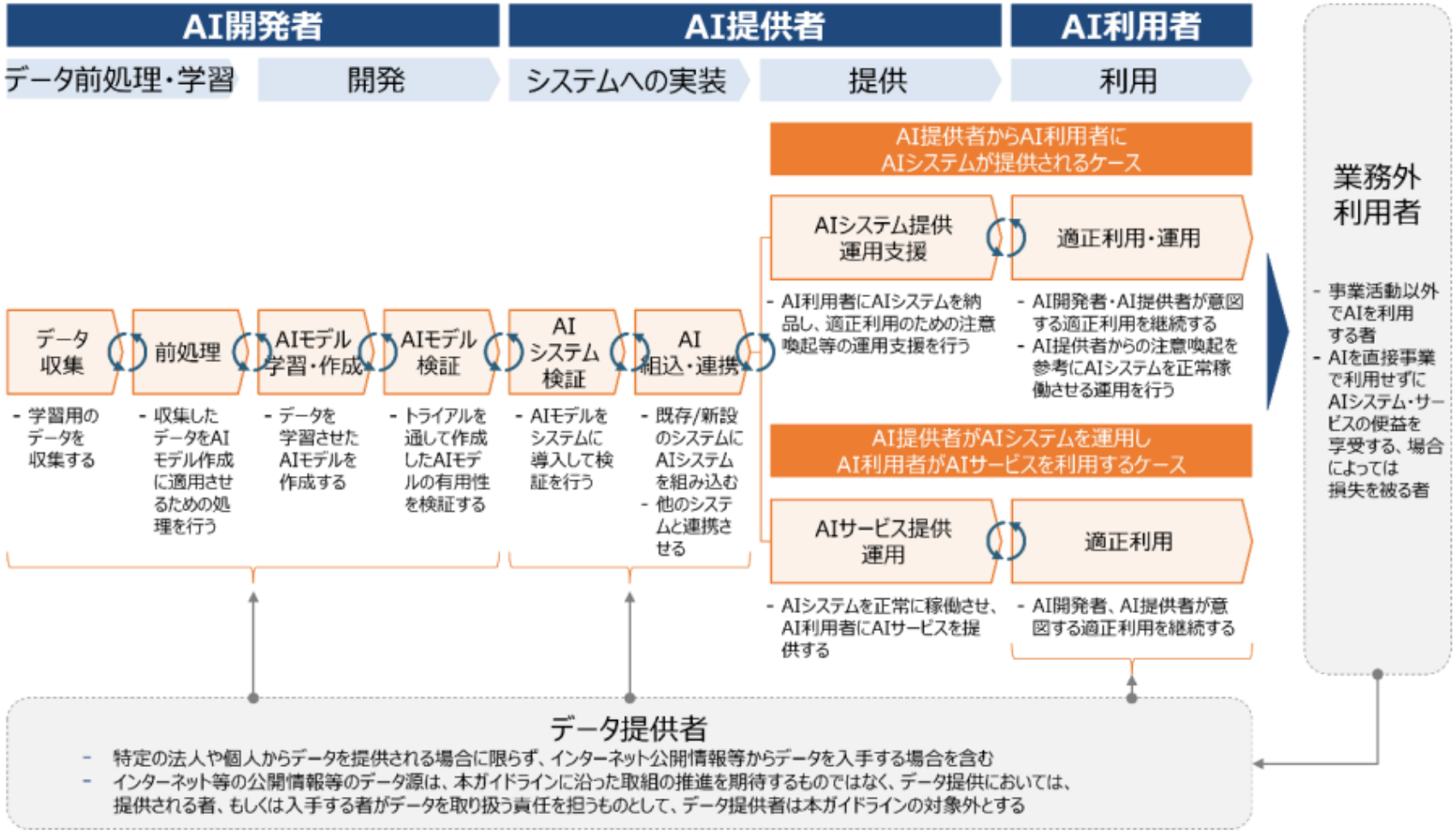
AISIでの実施予定事項２

- 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

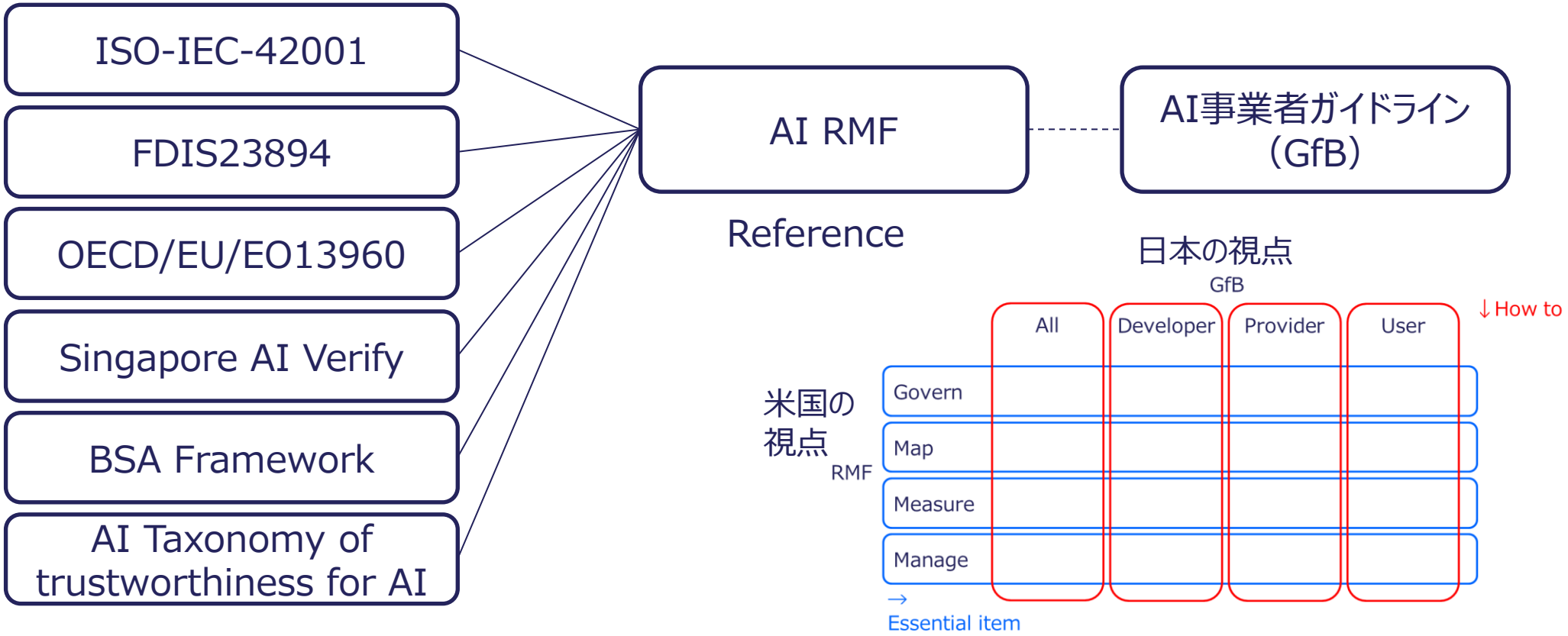
AI事業者ガイドラインの概要

- 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



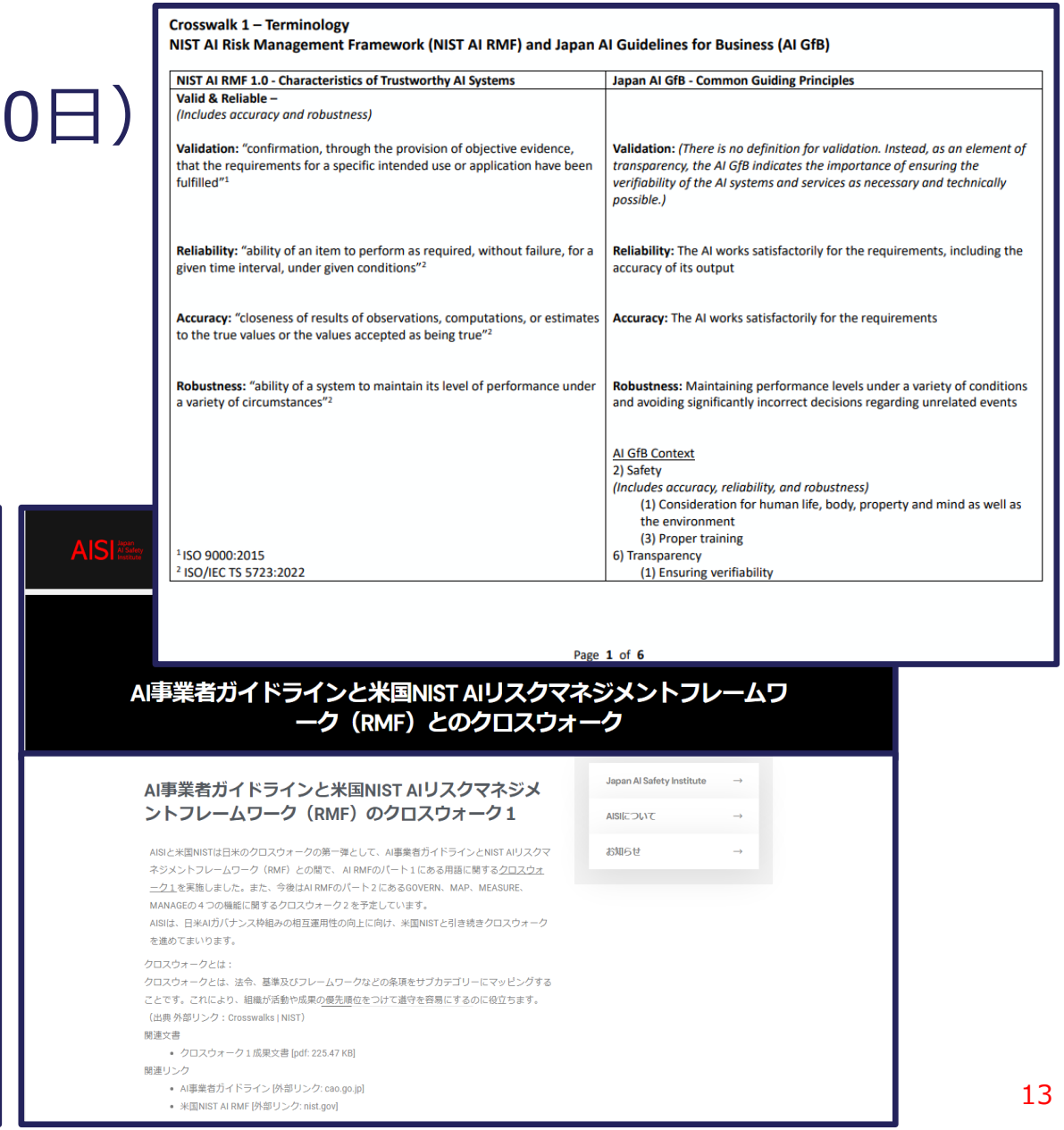
日米クロスウォークの概要

- 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キックオフ（5月1日）
  - 8月に成果公開予定


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

国際連携

AISI Japan AI Safety Institute

- AISI設立後、各国との意見交換を積極的に実施
  - 米国
    - スタンフォード大学でのイベントに登壇
      - 米国AISIケリー所長・英国AISIオリバー氏とパネルディスカッション
    - ケリー所長と「AISI間のグローバルネットワーク」の構築について意見交換
  - 英国、EU、シンガポール、オーストラリア、韓国は事務レベルの打ち合わせを実施。
  - AI関連事業者及び団体との打ち合わせを実施。

- AIセーフティサミット（5/21-22 韓国）
  - セッションへの参加、英米AISI等とのバイ会談

調査事業の方針

- 全体像の検討と対象領域の明確化
  - 関連する国内外の代表的なフレームワーク調査
  - 要素ごとの比較・重みづけ
  - 重点要素・領域の決定
- 評価手法の検討
  - 国内外のフレームワーク等から、評価に関する項目の抽出
  - AIシステムの評価サービスを提供する事業者に関する調査
  - AIセーフティに関する評価の考え方を示すガイド案の検討
- レッドチーミング手法の検討
  - 国内外のフレームワーク等から、レッドチーミングに関する項目の抽出
  - レッドチーミングを提供する事業者に関する調査
  - レッドチーミングに関するガイド案の検討

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。

AISI

Japan AI Safety Institute