AIセーフティに関する取り組みについて

2024-06-07

AIセーフティ・インスティテュート
副所長、事務局長
平本　健二

AISI Japan AI Safety Institute

IPA

Japan AI Safety Institute

AISI

AIリスクならびにAIを取り巻く世界動向

AIのリスクとは

AISI Japan AI Safety Institute
IPA

- **生成AIなどの台頭により顕著となったAIのリスク**
  - AIリスクとは以下の２つに大別
    - 技術的リスク（誤判定、バイアス、ハルシネーション、安全性、セキュリティ等）
    - 社会的リスク（プライバシー侵害、政治活動への悪用、不正目的、権力集中、財産権の侵害、環境負荷、心理的影響等）
  - 企業のAI活用には、さらに、法的なリスク、レピュテーションのリスクも存在
- **AI原則からAIガバナンスへ**
  - 2010年代にAIが浸透しはじめたときに、世界各国で作られたAI原則
    - 安全性・セキュリティ・プライバシー・公平性・透明性あるいは説明可用性
  - しかし、生成AIの出現により、透明性・説明可用性が困難に

AIの安全性の担保のためにはAIガバナンスが必要

AIガバナンスとはAIのリスクを受容可能な最小限に抑えつつ、
AIがもたらす価値を最大化することを目的とする

AIリスクの重要な視点

AISI Japan AI Safety Institute
IPA

1. AIリスクの状況は常にアップデートされる

2. AIガバナンスの目的は、リスクをゼロにすることではない

3. あらゆる組織、個人がAIリスクと対面する

4. AIリスクは提供する側・開発する側だけに関係するものではない

AIにおけるグローバルな枠組み

AISI Japan AI Safety Institute
IPA

♦ AI分野での世界における日本のリーダーシップ

ＡＩに関する国際ルール作りでのリーダーシップを発揮

- G7
  - 2016年のG7情報通信大臣会合(@高松)でAIに関する国際的な議論を提案。広島ＡＩプロセスを主導。
- OECD
  - エキスパート・ネットワークに参加し、OECD原則の策定(2019年)・改定(2024年)を含めた活動に貢献。
- GPAI
  - 創設に参加し、ステアリングメンバー国として推進。2022年にGPAIサミットを東京で開催。
- 国連
  - AIハイレベル諮問機関に構成員を推薦し貢献。

今後、技術・実装レベルでの枠組み整備でも積極的に貢献していくためのポジションを確保していくことが重要

♦ 安全性における国際的な連携も進行

- 英国、米国に設立されたAISIおよびAI安全関連機関との連携（後述）
- 5月21日、AIソウル・サミット(首脳セッション)のソウル宣言で、各国のAISI関連機関間のネットワークの育成により、安全性研究に係る協力を促進し、ベストプラクティスを共有する旨記載。

日本におけるAISIの設立

AISI Japan AI Safety Institute
IPA

- 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
- 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」に合意
  - 岸田総理大臣がAISI設立を表明
- 2024年2月14日
  - AIセーフティ・インスティテュート（AISI）を設立

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/＞

所長
村上　明子

各国のAI安全性確保への取組み

AISI Japan AI Safety Institute
IPA

- **米国**
  - AISIを設立
  - 基本は民間主導、民間とのコンソーシアムとの協働を強力に推進
  - 規模は30人程度。80名位を目指し推進中
- **英国**
  - AISIを設立
  - 政府主導、評価、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**
  - AI利活用も担当するAIオフィスで、安全性も推進。AI法の整備と推進も担う
  - 30人程度の規模
- **カナダ**
  - 国内機関の協力のもとAISI設立を準備中
- **シンガポール**
  - AISI相当機能を持つIMDA（情報通信メディア開発庁）がAI評価のためのAI Verifyを提供
- **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

統合イノベーション戦略2024（6月4日閣議決定）

統合イノベーション戦略における３つの強化方策

AISI Japan AI Safety Institute
IPA

重要技術に関する統合的な戦略

- 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）
- 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）
- 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）

グローバルな視点での連携強化

- 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）
- 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）
- 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）

AI分野の競争力強化と安全・安心の確保

- 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）
- **AI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）**
- 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分野の競争力強化と安全・安心の確保

AISI Japan AI Safety Institute
IPA

- 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

参考：② AIの安全・安心の確保

AISI Japan AI Safety Institute
IPA

イノベーション推進のためにもガードレールとなるAI利用の安全・安心を確保するためのル ールが必要である。我が国は、変化に迅速かつ柔軟に対応するため、「AI事業者ガイドライン」 に基づく事業者等の自発的な取組を基本としている。今後、AIに関する様々なリスクや、規格 やガイドライン等のソフトローと法律・基準等のハードローに関する国際的な動向等も踏まえ、 制度の在り方について検討する。

- 自発的ガバナンスと制度の検討
  - 幅広い業種に「AI事業者ガイドライン」の周知・浸透を図る。
  - 2024 年５月のAI戦略会議で了承された「AI制度に関する考え方」等を踏まえ、今夏にAI 戦略会議の下で新たに開催するAI制度研究会（仮称）において、制度の在り方の検討に着手す る。
  - 医療、自動運転、金融等の社会への影響が大きい重要分野は、技術の進展や利用状況に応じて制度の見直しの必要性等を検討する。
- AIの安全性の検討
  - AISIは、AIの安全性の中心的機関としてIPAに設置され、 AISIにおける専門人材の育成・確保、先進的な技術的知見の集約等を進める。関係省庁・機関等は内閣府が事務局を務めるAISI関 係府省庁等連絡会議を通じAIの安全性確保に向けた政府方針等をAISIが設置したAISI運営委員会に対して示すとともに、事業方針や計画、成果等について報告を受け、 AISIと協力する。
  - 外部知識を利用してハルシネーションを防止する技術などAIの安全性に関する最先端の研究開発を官民が連携して進める。
- 偽・誤情報への対策
  - 生成AIを利用したものを含め、ネット上に流通・拡散する偽・誤情報や、ＳＮＳ上のなりす まし型偽広告への対応等について、国際的な動向を踏まえつつ、技術・研究開発の推進、ファク トチェックの推進、国際的な連携強化など、制度面も含む総合的な対策を進める。
  - ネット上に流通するAI生成コンテンツを判別する技術の開発・実証等や、リテラシー向上等に取り組む。
- 知的財産権等
  - 内閣府「AI時代の知的財産権検討会」の「中間とりまとめ」や文化審議会著作権分科会法制度小委員会の「AIと著作権に関する考え方について」を踏まえ、今後の技術発展や海外動向等も見ながら、俳優や声優等の肖像や声も含め引き続き必要な検討を進めていく。

Japan AI Safety Institute

AISI

AISIと関連取り組みについて

（エイシーと読みます）

AISIの概要

AISI Japan AI Safety Institute
IPA

- **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
IPA

- **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く、関連の研究機関との連携を行う
- **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
IPA

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

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
- **各国との意見交換**

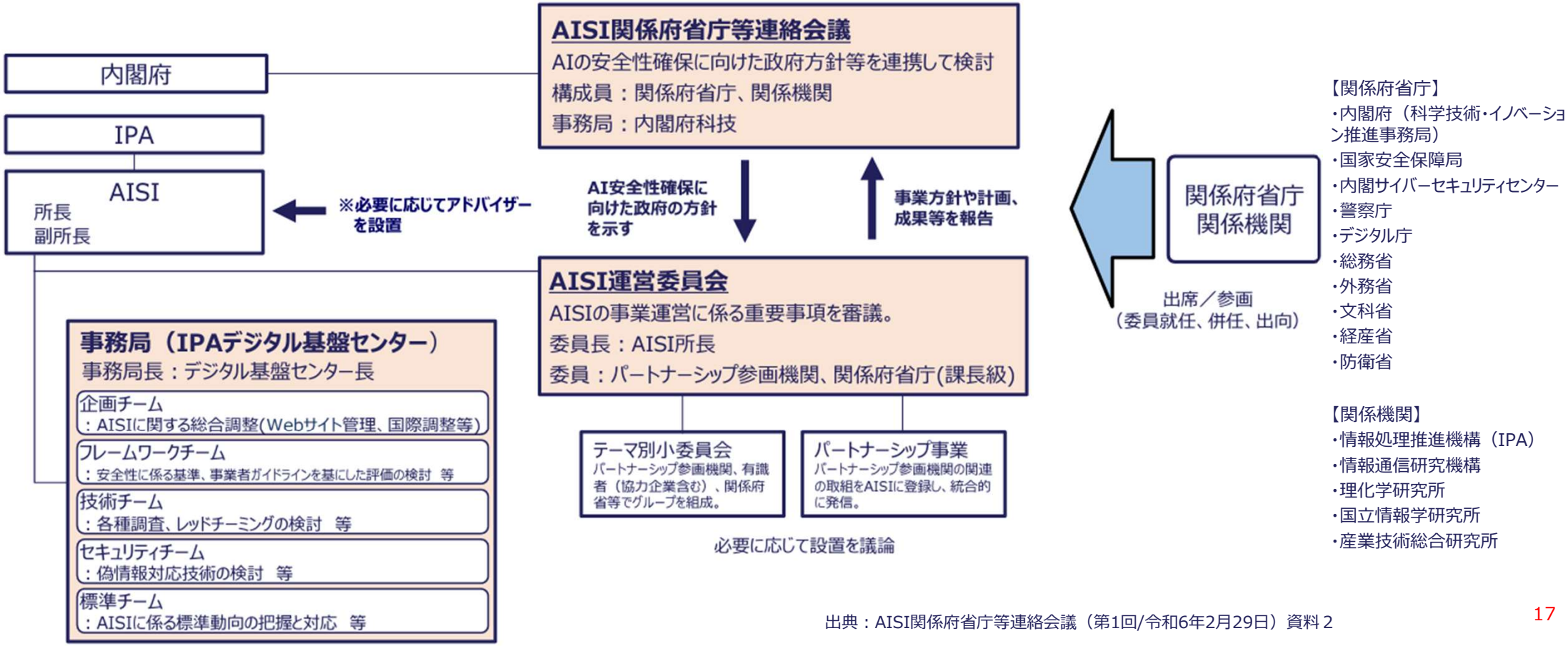
  AI関連事業者及び団体との事務レベルの意見交換を積極的に実施
  - 米国、英国、EU、シンガポール、オーストラリア、韓国との意見交換
  - 事業者等のエグゼクティブとの意見交換
  - GPAIワークショップ（パリ）参加（事務局、5月22・23日）

スタンフォード大学でのAIシンポジウム

AIソウルサミット同時開催のグローバルフォーラム

AISIの推進体制

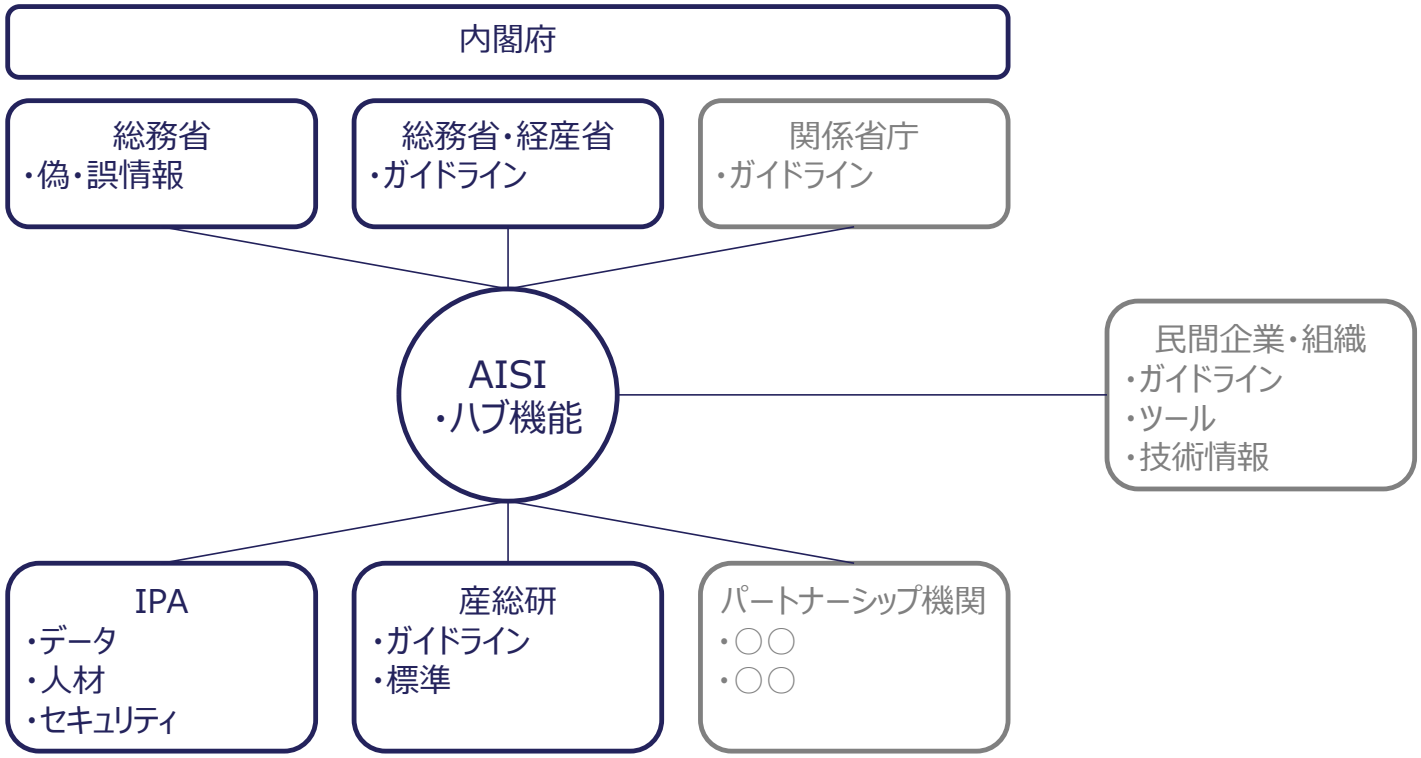
- 内閣府を事務局とする「AISI関係府省庁等連絡会議」で重要事項を審議。AISIに、AISI所長を委員長とする「AISI運営委員会」を設置し、その下に、必要に応じて、「テーマ別小委員会」や「パートナーシップ事業」を設置。



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

AISI推進における関係機関との協力関係

- 各機関の特徴を活かし、各機関の主体性と組織ブランドを活かしながら連携を図っていく。

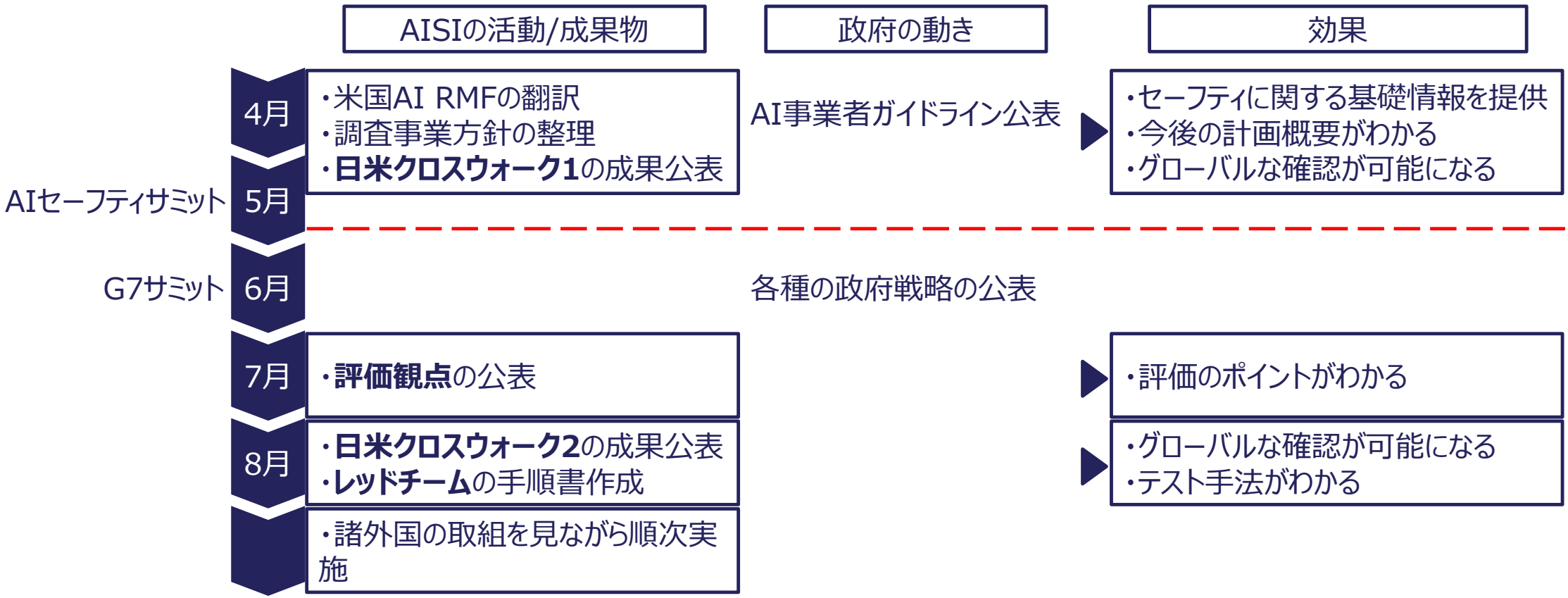


Japan AI Safety Institute

AISI

AISI関連活動のロードマップと成果実現に向けた直近の取組

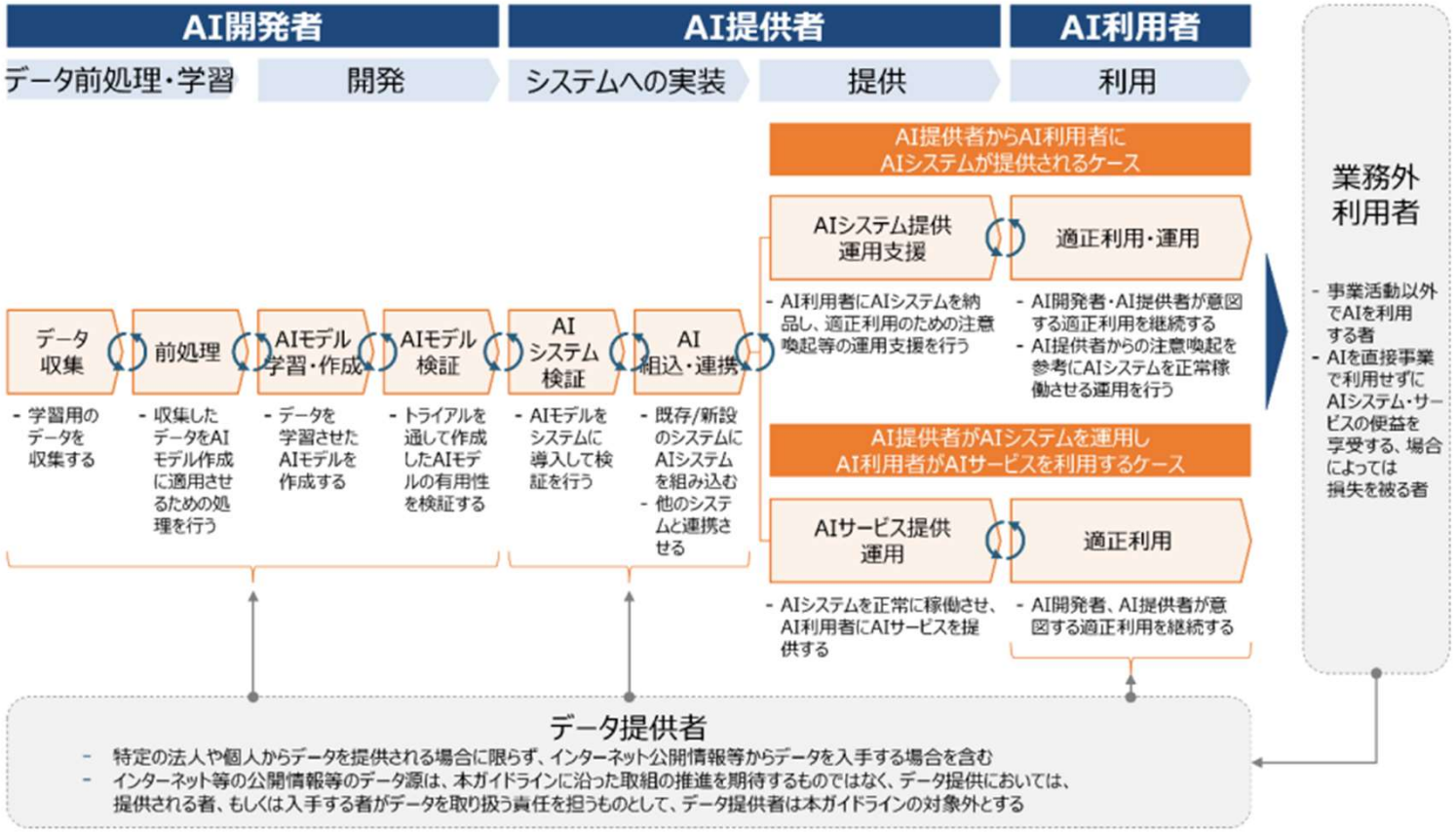
AISI 当面の活動と成果予定物



AI事業者ガイドライン（総務省、経済産業省）の概要

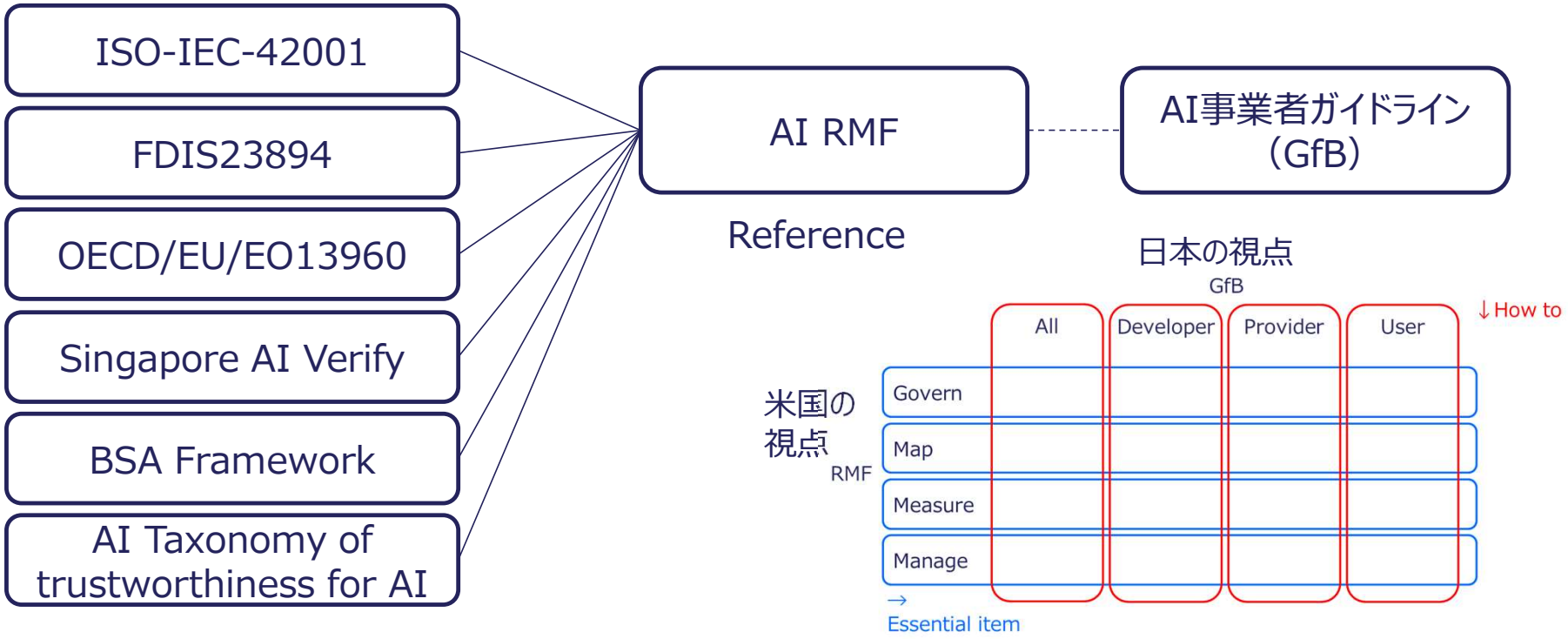
AISI Japan AI Safety Institute
IPA

- 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



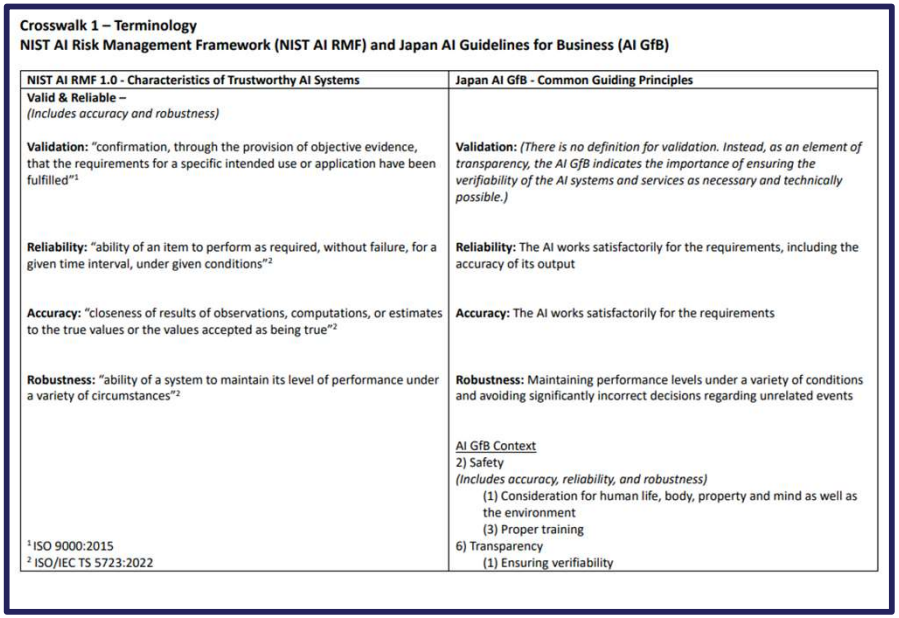
日米ガイドラインのクロスウォークの概要

- 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）
  - ターミノロジ
- クロスウォーク２を実施中
  - 項目レベル（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

関連調査

AISI Japan AI Safety Institute
IPA

- AIセーフティ関連調査（公開中）
  - 米国におけるAI政策最新動向調査
    - https://www.ipa.go.jp/digital/chousa/trend/m42obm000000ic9a-att/ny-dayori202405.pdf
  - 米国におけるAIのセキュリティ脅威・リスクの認知調査
    - https://www.ipa.go.jp/security/reports/technicalwatch/20240530.html
- 評価手法、レッドチーミング関連調査（実施中）
  - 評価手法、レッドチーミング関連ドキュメント作成
- 国内関連団体調査（実施中）
  - コミュニティー形成やコラボレーションのための基礎調査
- テストベッド（検討中）
  - 国内外の事例調査含め、検討中
- 標準
  - ISO/IEC SC42等の標準化推進状況等の調査

AISI Japan AI Safety Institute

AI安全性への取り組みと今後について

AI安全性の現在と今後の動き

AISI Japan AI Safety Institute
IPA

- **AISI設立後の活動は順調**
  - 欧米との連携も進みじはじめ、政府内の連携もチームメーキングが進んできている。
- **今後のAISIの活動**
  - 具体的なアウトプットを基に、国内展開を図っていく。同時に、グローバルなポジショニングを確保していく。
  - 民間の取り組みとの協力を強化していく。

人材を大募集中です！

世界最先端のグローバルな枠組みつくりに挑戦したい人は、ぜひ一緒に取り組みましょう！！

AISI

Japan AI Safety Institute

IPA