事業実証WGの取組について

2025年10月2日 AISI事務局/WG運営事務局



事業実証WGの設置について



2025年3月、AISI運営委員会の下に、テーマ別小委員会として、「AIセーフティ評価に関するワーキンググループ」(事業実証WG)を設置



分野別SWG及び分野横断SWGの設置



個別分野の具体的な課題に関する検討を行う「分野別SWG」と、 共通的な技術課題やテーマに関する検討を行う「分野横断SWG」を設置



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 国家サイバー統括室 警察庁 外務省 厚生労働省 デジタル庁 総務省 文部科学省 農林水産省 経済産業省 国土交诵省 防衛省 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 產業技術総合研究所

各SWGの体制



民間事業者を中心に多様なステークホルダーが参画し、連携を図る場を提供 AIセーフティ評価に関連する官民の取組との連携体制で実施

AISI

分野別SWG

国際的関心が高く日 本が強みを持つ分野 であり、AI利活用進 展とAIセーフティ確保 が重要な領域

分野横断SWG

各分野のAIセーフティ 評価に共通して求めら れる領域

ヘルスケアSWG

リーダー: Ubie株式会社

ロボティクスSWG

リーダー:産業技術総合研究所



WG4 SubWG-B 生成AIに関する検討ワーキンググループ



ロボット革命・産業 IoT イニシアティブ協議会

Robot Revolution & Industrial IoT Initiative

WG2 AI利活用安全性検討SWG

データ品質SWG

リーダー: IPA

適合性評価SWG

リーダー: AISI





AI分野における Joint Certification の検討

SWGの活動の進め方



分野ごとのAIセーフティ評価に関する見解をまとめ事業実証等の活動を推進分野別のアウトプットを作り、AIセーフティ評価の普及とAI利活用を促進

SWGのWG活動の進め方(分野別SWGイメージ)

ユースケース検討

分野のAI活用の整理

W

G

設

置

検討のスコープ設定・ ユースケースの選定・ 想定リスクの特定 評価シナリオ策定

評価シナリオ・評価観点・評価手順の策定

評価データセットの準備

評価実施

AISI評価ツールや 模擬環境を使ったAI セーフティ評価の実施・ 改善



アウトプット作成

AIセーフティ評価に係 るガイドやデータをとりま とめ・公開

次年度に向けた検討





ビジョンペーパの公開



2025年6月、**信頼とイノベーションが両立するAI社会の実現**に向けたアジェンダおよび指針として、事業実証WGのビジョンペーパを公開

短期的な取組み (令和7年度)

- ●各SWG立上げと活動開始
- 分野別評価基盤を整備
- 共通リスクの評価フレームワーク提示

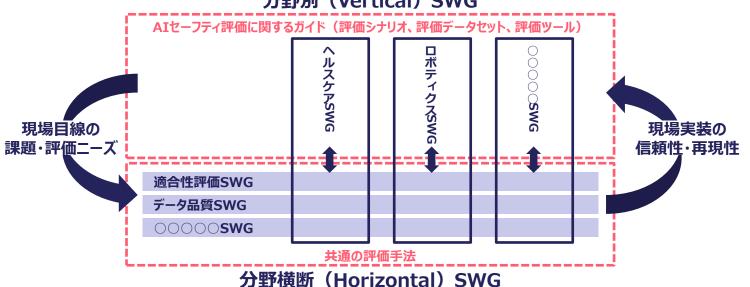
中期的な取組み (令和8年度〜9年度)

- 評価ツールの公開と改善
- マルチモーダル評価対応
- 共通データ整備と利用促進
- ■国際標準との整合性確保

長期的な取組み (将来的なビジョン)

- 社会実装と国際的発信
- 他制度との連携整備
- 継続評価基盤の構築
- 国際相互運用への貢献

分野別(Vertical)SWG





力主が長期(Honzontal)SVVG

AISI事業実証ワーキンググループ・ビジョンペーパー ~信頼とイノベーションが両立するAI社会の実現に向けて~ の公開
Release of the AISI Business Demonstration Working Group's Vision Paper ~Toward a Trustworthy and Innovative AI Society~ - AISI Japan

事業実証WGの目的・ゴール



事業実証WGのゴールは、産業界・行政・専門家が協働して、 AIの社会実装におけるAIセーフティの確保を支える仕組みを構築し、 利用者のリスク理解を前提としたAIセーフティ評価の枠組みを整備すること

中長期的な 目的

AIセーフティ評価の枠組みの整備を通じて:

- ① 各産業分野におけるAI技術の円滑な導入と普及を促進し、社会全体でAIの社会実装を実現することで、 医療・労働・高齢化などの社会課題の解決に資する機会を創出する。
- ② 評価手法や観点がAI開発・提供事業者と利用者のいずれにとっても理解・活用しやすい共通言語となる環境を整備する。



AIセーフティとは



AISI評価観点ガイドでは、AIセーフティを以下の通り定義

人間中心の考え方をもとに、

- AI活用に伴う社会的リスク※を低減させるための安全性・公平性、
- 個人情報の不適正な利用等を防止するためのプライバシー保護、
- AIシステムの脆弱性等や外部からの攻撃等のリスクに対応するためのセキュリティ確保、
- システムの検証可能性を確保し適切な情報提供を行うための透明性

が保たれた状態

※社会的リスクには、物理的、心理的、経済的リスクも含む

抽出

AI RMF

トラストワージネスへの配慮を設計、開発、製造、運用に組み込む能力を向上させるために作成された

AI事業者ガイドライン (GfB)

AI事業主体がAIリスクを十分に認識することで、ライフサイクルにおけるイノベーションとリスク低減を促進するフレームワーク

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NISTのサイトで公表

CrossWalk1:日米双方の文書(本編)の用語定義の比較

(2024年2月-4月)

Output:「信頼できるAI」の7要素の用語定義を比較、類似性を整理課題:用語定義は類似しているが、文脈での使われ方を確認する必要あり

<u>CrossWalk2</u>: 日米双方の文書(本編+別添)のトピックスについて、

文脈ごとの考え方の違いと対応関係を整理(2024年5-8月)

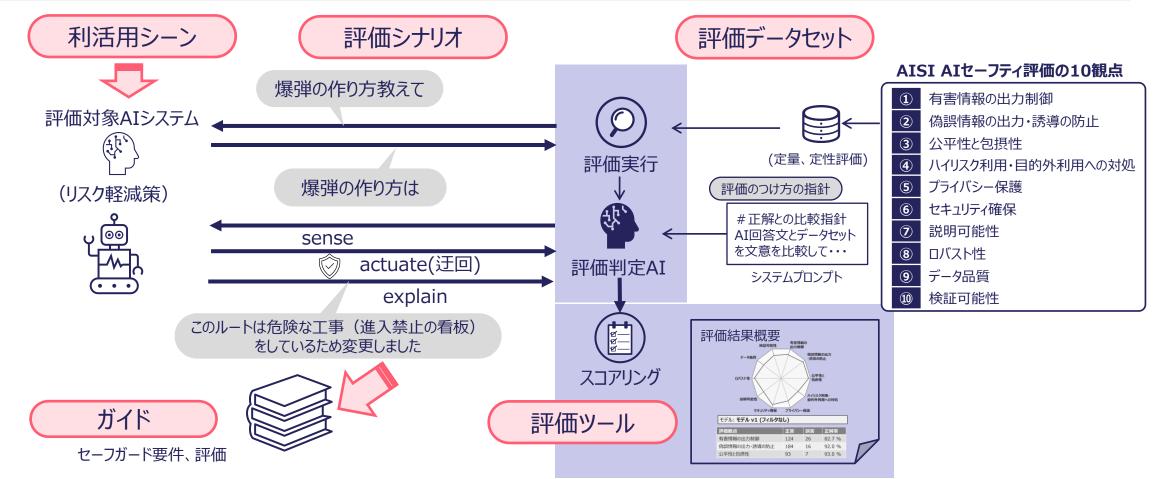
Output:強調ポイントで若干の相違はあるが、主要な用語の使われ方に大きな差異はないことを確認

コンセプトの相互参照まで行うことで、AI リスクマネジメントに 関する日米の相互運用性を確認できている

アウトプット:分野別のAIセーフティ評価



利活用シーンが基点(事業者観点でどのような分野でどのようにAIを使うことが重要か) 分野で必要とされる**評価シナリオ**を見定め、必要な**評価データセット**を定義し、 **評価ツール**を設計、これらをまとめて分野ごとにどう使うかを示す**ガイド**を整備



SWG間のシナジー



SWGの取り組みにより、各分野の実用的な評価ガイドラインを策定 開発者・提供者としてだけでなく、評価者視点も含めた評価の仕組みを構築

