

ビジョンペーパー

AI利活用促進に向けたAIセーフティ評価に関する事業実証

～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～

2025年6月30日

AI セーフティ・インスティテュート
事業実証ワーキンググループ

AISI Japan
AI Safety Institute

ビジョンペーパーの位置づけ

- 本ビジョンペーパーは、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、社会・産業・政策の各レベルにおける**AIセーフティ評価に関する共通理解の醸成と具体的な実装への道筋を示す**。

AIセーフティの 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

AISI『AIセーフティに関する評価観点ガイド』より

本ビジョンペーパー の 想定読者

AI開発・提供事業者



基盤モデルやプロダクト・サービスに関する
AIセーフティ評価を行う場合の参考として参照

AIの導入・利活用を行う企業・自治体等



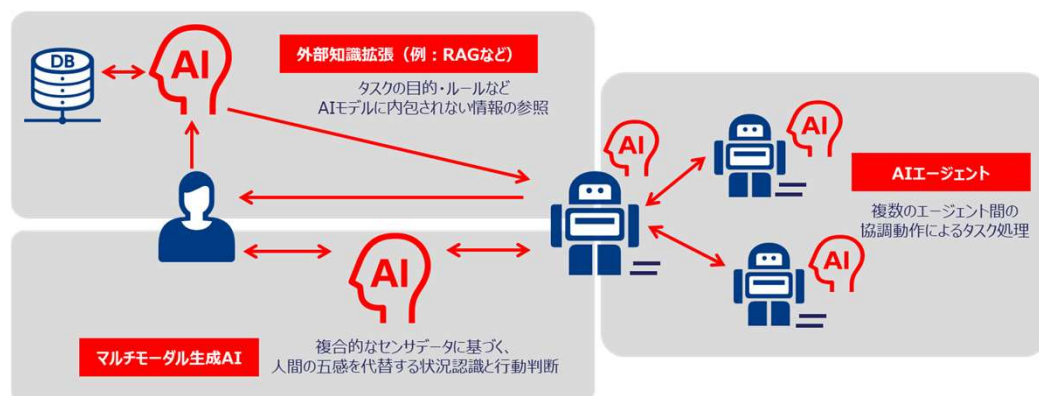
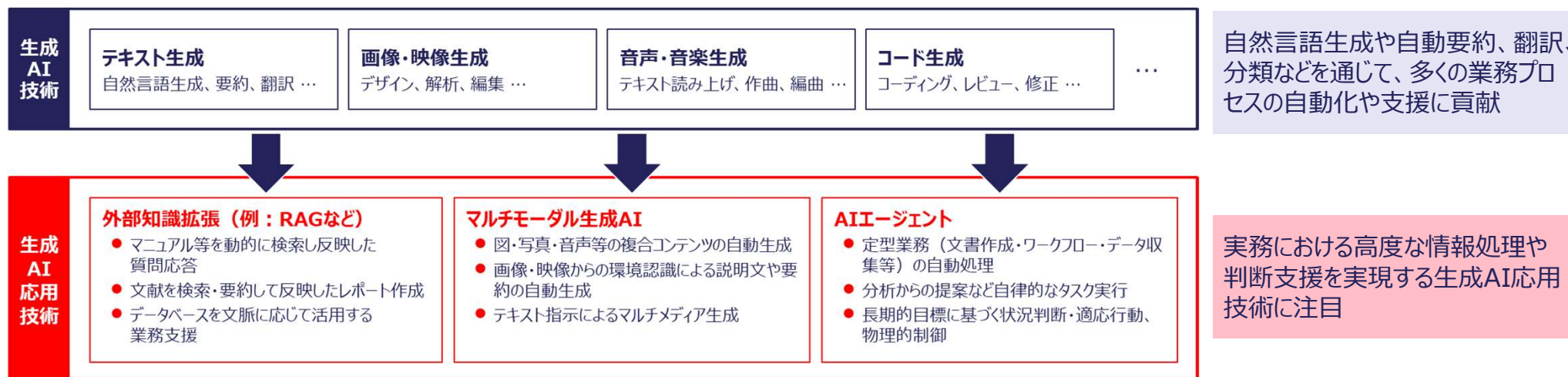
AIセーフティに関する課題整理や受容性の判断に
資する観点を把握する上での手がかりとして活用

政策担当者・認証機関等



AIセーフティに関する今後の評価指針の
策定等に向けた議論の参考として参照

- 生成AIを基盤とする応用技術の進展はAIの社会的役割やシステム内での位置づけに質的な変化をもたらしている。



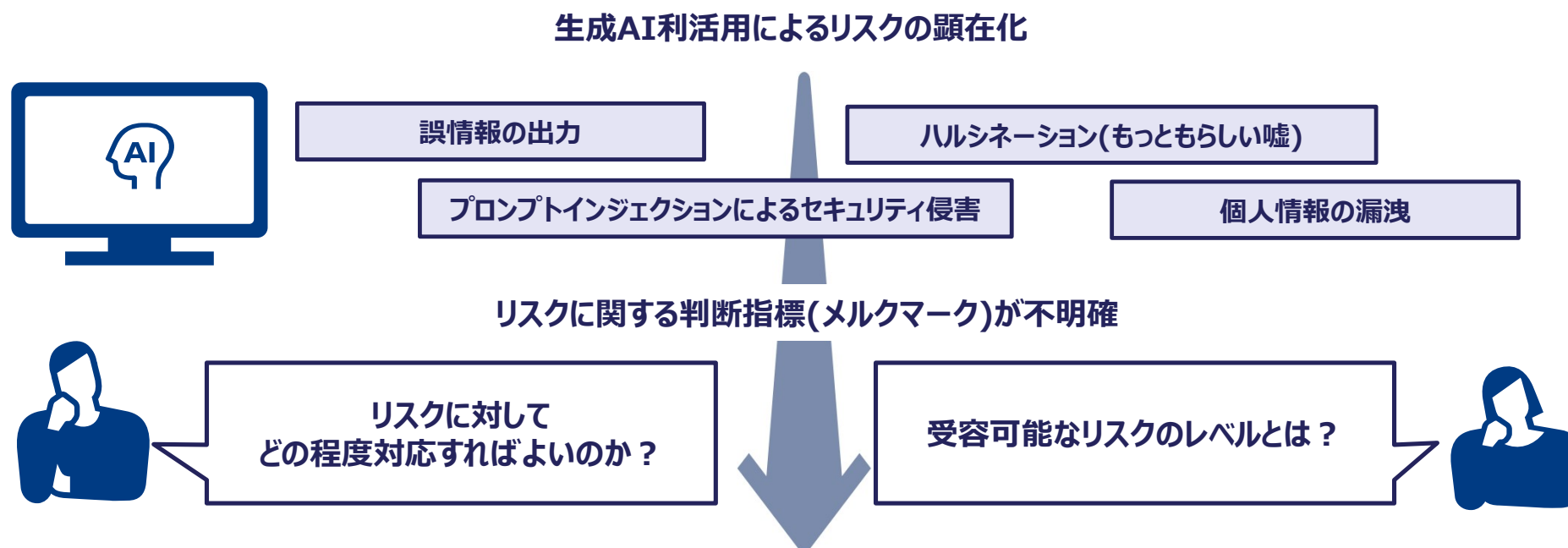
生成AI応用技術により、デジタル空間での応用に加え、物理空間での行動制御や運用支援への展開も期待

実装例：

- 災害現場で探索を担うドローン
- 工場内で複数台で協調して搬送を行うロボット

AIが社会インフラとして機能する新たな段階へ

- AI利活用が急速に拡大する中、新たな便益がもたらされる一方で、**社会的信頼を損なうリスク事象も数多く顕在化**。
- AIの用途やリスク特性に応じたセーフティ評価フレームワークを整備することが急務。



AIの用途やリスク特性に応じたAIセーフティ評価の枠組みを整備する必要性

適切なAIセーフティ評価により、AIが社会において信頼に足る存在（「信頼に足るAI」）であることを示すとともに、「信頼できるAI」として受け入れられるための基盤を形成

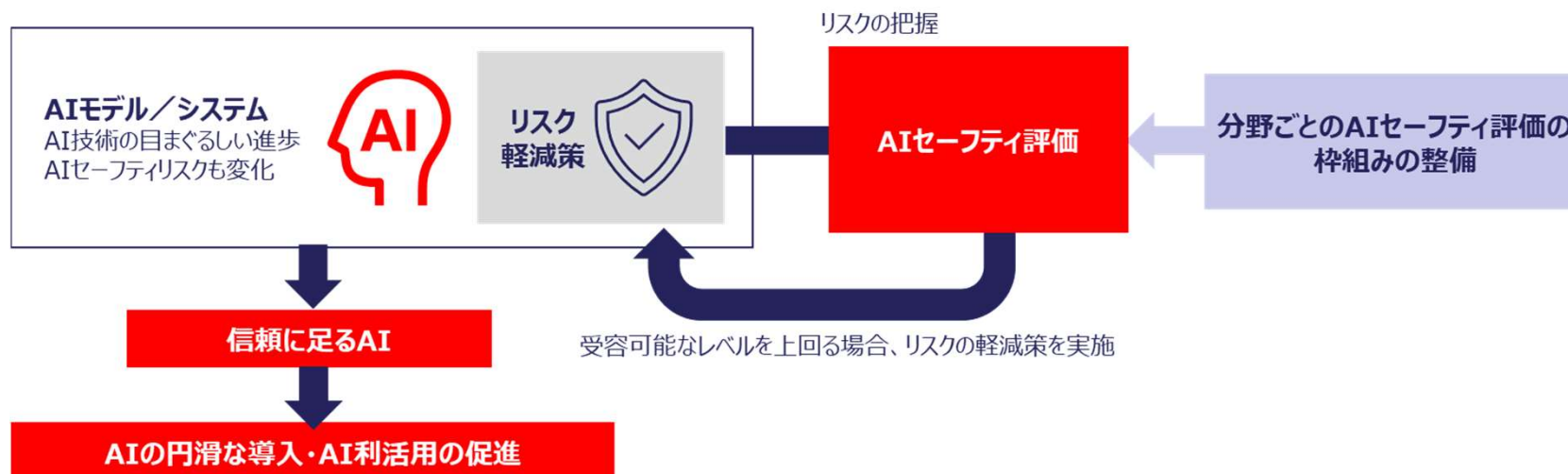
事業実証WG*の目的・ゴール

- 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：

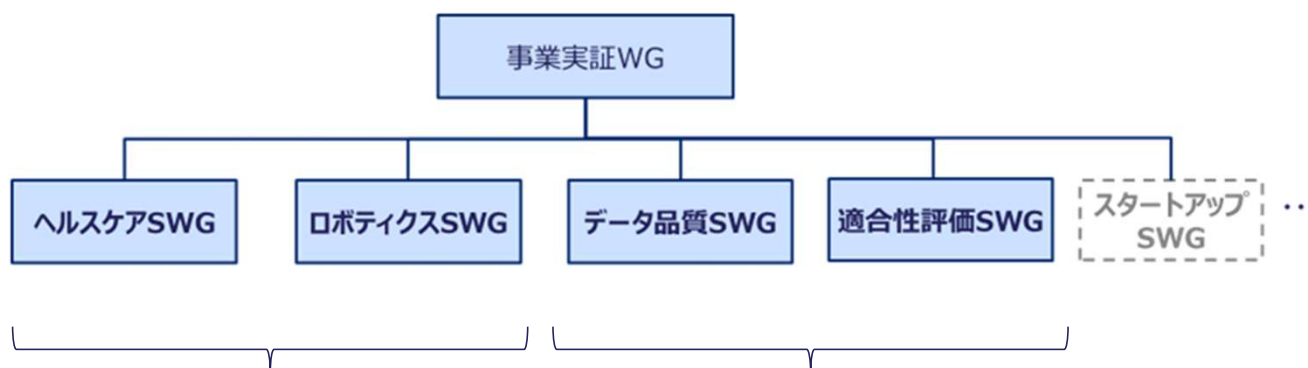
- ① 各産業分野におけるAI技術の円滑な導入と普及を促進し、社会全体でAIの社会実装を実現することで、医療・労働・高齢化などの社会課題の解決に資する機会を創出する。
- ② 評価手法や観点がAI開発・提供事業者と利用者のいずれにとっても理解・活用しやすい共通言語となる環境を整備する。



事業実証WGの実施体制

- **分野別（Vertical）SWG*と分野横断（Horizontal）SWGを組み合わせた体制を構築。**
- 令和7年度は、4つのSWG（ヘルスケア、ロボティクス、データ品質、適合性評価）で構成。

*SWG：サブワーキンググループ

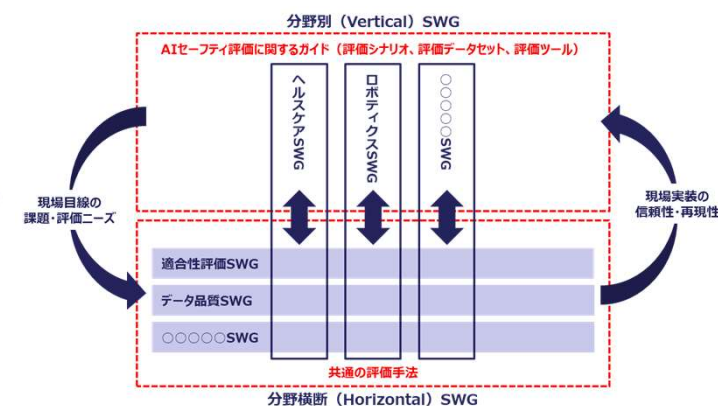


分野別（Vertical）SWG： 分野向けの評価ガイド等を検討

- 国際的関心が高く日本が強みを持つ分野
- 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（Horizontal）SWG： 分野共通の評価手法を検討

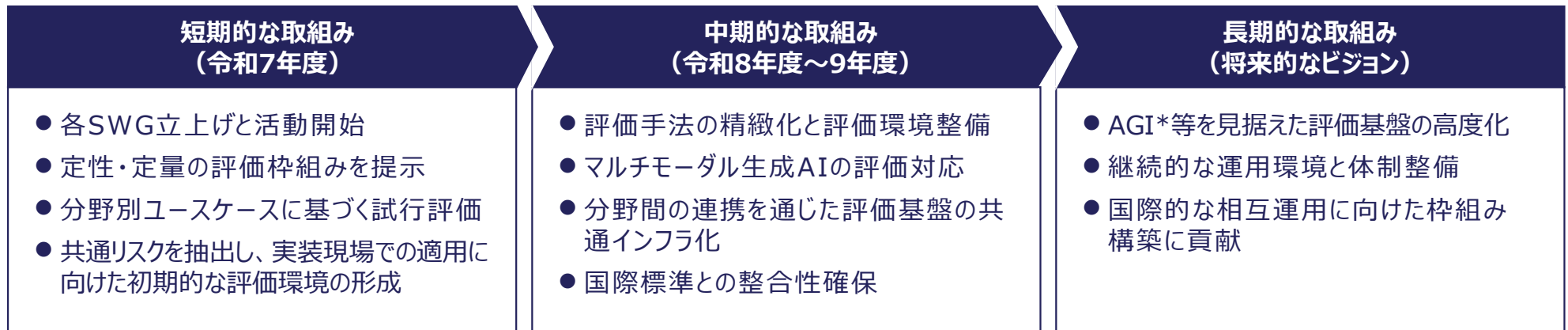
- 日本発の実践／重層的な評価枠組みの構築
- 国際的・標準的なルール形成に寄与



- 評価観点やツール整備、他SWGとの連携を進める。必要に応じてSWGの増設も計画する。
- Vertical／Horizontal SWGは、それぞれの役割を果たしつつ、相互補完・連携する。

事業実証WGのロードマップ

- AIの社会実装を促進するAIセーフティ評価の確立に向けて、段階的・継続的な取組みを実施。
- 令和7年度は重点分野のユースケースを通じて初期的な評価基盤を形成し、中長期的には技術多様性に対応した評価手法や環境を整備、国際標準との整合を図る。



*汎用人工知能（Artificial General Intelligence : AGI）

ヘルスケアSWGの概要（1 / 2）

- 医療現場の業務負担軽減や一般生活者の健康管理支援に生成AIの活用が進む一方、情報管理や誤出力への懸念から、活用ガイドの策定が進められている。
- 生成AIの情報が利用者にどう影響するかを予見し、制御・検証可能にする評価の枠組みの整備が必要。

ヘルスケア分野のAI利活用の現状

- ・医療機関における時間外労働条件規制
⇒医療従事者の書類作成業務などの負担軽減
- ・一般生活者の健康管理の商品・サービスも普及
- ・分野独自の取組み：
ヘルスケア事業者のための生成AI活用ガイド*策定

AIセーフティ評価の重要性と利活用促進への期待

- ・ヘルスケア分野は生命・身体に影響可能性あり
⇒ハルシネーションや情報管理に懸念
- ・JaDHAガイドやメルクマールが一定程度可視化
⇒評価項目や方法は未整備
- ・ユースケース毎の評価手法・検証事例の体系化
⇒安心・安全な利活用環境の整備

医療現場における生成AI活用の例（退院時サマリ作成等）

*日本デジタルヘルス・アライアンス(JaDHA)『ヘルスケア事業者のための生成AI活用ガイド』（2025年2月第2.0版公表）



既存のガイド等を踏まえ、医療現場の実務ニーズとリスク懸念を起点としたAIセーフティ評価の枠組みの整備を進める。

ヘルスケアSWGの概要（2 / 2）

- 生成AIのヘルスケア分野での利活用には、プライバシー漏洩や誤情報、バイアスなど多様なリスクがあり、ユースケースごとの評価観点と定量・定性的な評価モデルの整備が重要である。
- チェックリストや評価ツール、実データによる検証を通じて、安全性評価の実効性向上を目指す。

想定される リスク	・ プライバシー漏洩 ⇒マスキング処理不完全 ・ ハルシネーション ⇒診療ミスや誤情報提供 ・ 説明可能性の欠如 ⇒出力根拠不明瞭、現場が理解困難 など	AIセーフティ評価 の方向性	・ 評価観点をユースケースごとに整備 ・ 定量・定性的な評価モデルの構築 ・ Non-SaMD*のテキスト生成AIを対象 ・ 以下により安全性評価の実効性を高める ✓ チェックリスト形式の評価項目集の作成 ✓ 評価ツールの開発 ✓ 実データを用いたフィジビリティ検証
SWGの ロードマップ	短期的な取組み (令和7年度)	中期的な取組み (令和8年度～9年度)	長期的な取組み (将来的なビジョン)
	● Non-SaMDを対象に検討対象とする 代表的ユースケースの選定 ● 生成AI利活用のリスク構造の明確化、 セーフティ評価の観点・シナリオの設計 ● 評価ガイド・データセット・ツール等検討	● AI開発・提供事業者による評価ガイ ド・データセット・ツール等の開発・効果 検証 例：複数の医療機関で導入される生 成AIプロダクト・サービスにおいて試行 的にAIセーフティ評価を実施	● 評価ガイド・データセット・ツール等の ヘルスケア業界での広報・浸透活動 ● 評価ガイド・データセット・ツール等の継 続的アップデート

*医療機器または医療機器プログラムには該当しないプロダクト・サービス（Non-Software as a Medical Device）

ロボティクスSWGの概要（1 / 2）

- ロボティクス分野では、「人とロボットのコミュニケーションやコラボレーション」を考慮した上で、AIセーフティを検討することが必要。
- AIとロボットの複合的な振る舞いが生む便益と、そこに内在するリスクのバランスを捉えた評価の枠組みの構築が重要。

ロボティクス分野のAI活用 の現状	・対人インターフェース領域におけるマルチモーダルAI導入による人間の五感的判断代替を模索 ・協働ロボットにおける人とロボットの協働／ロボット同士の協働 ・「Embodied AI」の進展による、状況理解と行動選択の高度な融合	AIセーフティ 評価の重要性 と活用促進 への期待	・対話型・移動型・協働型ロボットにおける対人リスクや状況対応の課題 ⇒機能安全では捉えきれない ・協調行動などのユースケースを想定し、「ロボットの種類×AI機能」のマトリクスで分類・評価する枠組みが必要
------------------------------	--	--	--

具体的ユースケースに基づく評価シナリオの評価方法の設計

	対話型	移動型	協働型
特徴	音声認識、音声合成、感情認識などの技術を統合し、自然言語による双方向のコミュニケーション能力を持つ。 利用者の意図や感情を理解し、適切な応答を生成することが可能である。	センサやカメラを用いて周囲環境を認識し、自己位置推定と経路計画による自律移動能力を持つ。 障害物回避や動的環境への適応が可能である。	人間と一定の距離を保ち、動作を監視しつつ、必要に応じて停止し、衝突を回避する。力加減や動作速度の調整による柔軟な作業が可能であり、人間との共同作業により、作業効率の向上が期待される。
ユースケース 分類（例）	● 接客支援 - 顧客への商品案内 - 自然対話による接客 等 ● 介護・メンタルケア - 音声・表情解析による感情推定 - 対話を通じた共感・励まし 等 ● 観光案内 - 外国人観光客向けの観光案内、経路案内 - おすすめスポットの紹介	● 自動配送 - 動的に経路を作成 - 渋滞等の環境変化を特定し、経路変更 等 ● 自動清掃 - 障害物回避 - 汚物等を検知し、適切な清掃方法を特定 等 ● 警備・見回り - 施設内の見回り・正常確認 - 異常検知した場合に駆け付け 等	● 作業指示の理解 - 人間からの口頭による指示を理解 - 指示された内容を行動に変換 等 ● 手順書・マニュアル作成 - 作業内容のモニタリング - 手順書やマニュアルの作成 等 ● 人手の動きの再現 - 熟練作業者による製作方法を詳細に認識 - 作業スタイルを再現 等

従来の機能安全に加え、信頼形成や文脈理解などを含むAIセーフティ評価の設計・検証の枠組みの整備を進める。

ロボティクスSWGの概要（2 / 2）

- ロボティクス分野のAIセーフティ評価では、物理的リスクに加え、心理的・社会的リスクの抽出と制御が不可欠である。
- 本SWGは、実環境と模擬シナリオを活用し、ロボット類型ごとの多層的評価を進め、将来的な標準的枠組みの確立を目指す。

想定される リスク

- 物理的リスク
⇒衝突・接触、感電・火災 等
- 心理的リスク
⇒恐怖・威圧的発言 等
- 社会的リスク
⇒差別的扱い、偽誤情報の拡散 等

AIセーフティ 評価の方向性

- AIロボットの便益と新たに生じるリスクの受容可能性を中心に検討
- 実環境 *での評価実証と模擬シナリオによるシミュレーションを組み合わせた評価設計



* ソーシャルイノベーション共創拠点「KAWARUBA」での実証を想定

SWGの ロードマップ

短期的な取組み (令和7年度)

- ロボット類型ごとのインタラクション起点のAIセーフティリスクを抽出
- 心理的影響・社会的受容性を含む多層リスク評価シナリオの開発
- 模擬環境でのリスクへの対応策の検討と初期的な試行評価

中期的な取組み (令和8年度～9年度)

- 実環境におけるロボット導入による効果検証
- リスクに関する指標の体系化
- 受容性・信頼性を含む統合的分析手法の確立・高度化

長期的な取組み (将来的なビジョン)

- 評価モデルや設計指針の体系化
- 幅広いロボット用途への適用を支える標準的枠組みの確立
- 運用段階での継続評価・改善を可能とする仕組みの整備

データ品質SWGの概要

- AIセーフティの基盤となるデータに注目して、データ品質の可視化と管理手法の確立を目指す。
- 既存の標準類を補完する形で実務的なガイドや評価ツールを整備し、分野別SWG等にて適用する。

現状と課題

- 入力データの欠陥が予期せぬ出力につながる
⇒データ品質可視化、管理手法の確立と実践
- 複数の国際標準、理論的な枠組み整備
⇒具体的な評価方法や導入手順は未整備
⇒多くの現場で実装上の障壁
- 『データ品質マネジメントガイドブックVer.1.0』公表
(AISI : 2025年3月)
- 必要十分でリーズナブルなデータ品質管理が重要

データ品質の確保に向けた取り組み方針

- ガイドの整備
 - ✓ 評価観点と実践フレームワーク整理
 - ✓ ガイドブックのアップデート
- 評価ツールの提供
 - ✓ マネジメントの第一歩として活用
 - ✓ 実践障壁の低減
- 適用検証とフィードバックの反映
 - ✓ 業務シナリオに即したデータ品質評価トライアル実施

SWGのロードマップ

短期的な取り組み (令和7年度)

- データ品質マネジメントガイドブック Ver.1.0に基づく適用検証の実施
- 簡易的な品質評価ツール（チェックリスト等）の試作と評価
- 試行的評価を通じた導入容易性・実効性の検証

中期的な取り組み (令和8年度～9年度)

- 複数分野での横断的導入実証による活用モデルの確立
- マルチモーダルデータやマルチエージェント等の新たな利用形態への対応
- 評価ツール機能の本格化

長期的な取り組み (将来的なビジョン)

- 組織内での継続的なデータ品質評価の運用モデル確立
- 技術進歩と社会変化に応じてガイドと評価ツールを更新する仕組みの構築
- 日本発のデータ品質評価フレームワークや指標定義の国際的な議論へ反映

適合性評価SWGの概要

- 分野別SWGの取組みを通じて、自己適合評価を含むAI適合性評価手法の確立を目指す。
- データ品質SWGとの連携により、AI出力の整合性、利用文脈を含めた複合的な評価の方法を検討する。

現状と課題

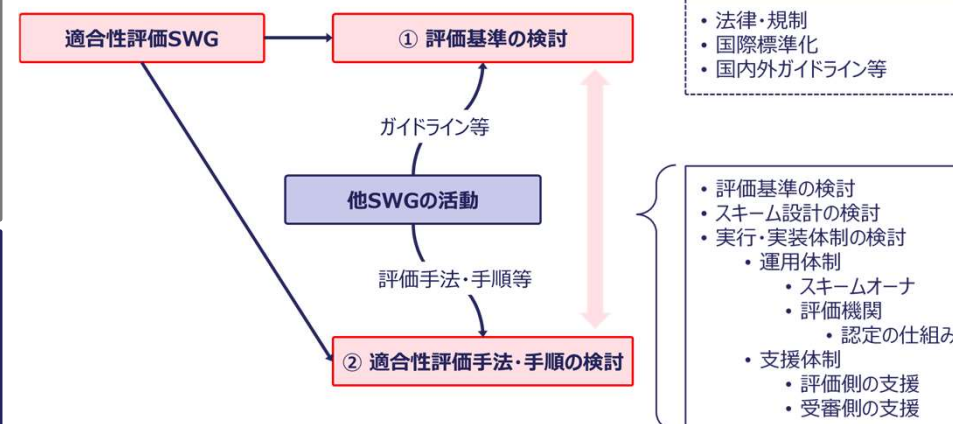
- 従来の適合性評価制度では、変化し続ける対象への評価に対応しきれない
- ISO/IEC 42007開発中(適合性評価)はハイレベル枠組み
⇒柔軟かつ適切な評価方法と実行可能手順が必要

適合性評価手法の確立への取組み方針

- 分野別SWGの取組みから、AIへの要求事項等導出
 <具体的なステップ>

- データ品質SWGと連携し、複合的な評価方法の検討

適合性評価手法の確立への取組み方針



SWGのロードマップ

短期的な取組み (令和7年度)

- 分野別SWGとの連携による評価ニーズや受審組織の現状に関するヒアリング
- 適合性評価の関連機関を中心としたSWGを組成し、既存の適合性評価手法を分析

中期的な取組み (令和8年度～9年度)

- 国際標準化議論を踏まえた、包括的なアシュアランスとアカウントビリティ確保等に係る適合性評価制度の設計
- AIに関する適合性評価制度のあり方の具体化と試行により実効性を検証

長期的な取組み (将来的なビジョン)

- 国際的な実装モデルとの整合性を踏まえたガイドライン整備
- 国際的に総合運用可能な実運用体制構築を視野に入れた段階的な展開

