

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～

2026年10月6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

- 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及び指針として、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を基盤とする応用技術の進展

RAG（検索拡張生成）

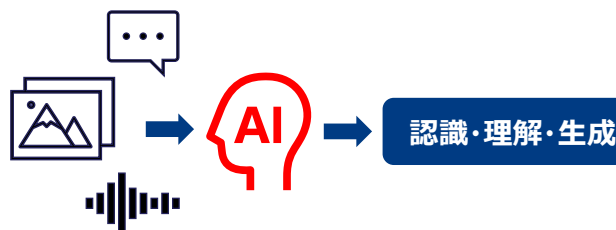
外部データベース等をリアルタイムに検索し、その結果に基づいて出力を生成



活用例：企業内部ナレッジを活用したFAQ応答・業務支援ツール

マルチモーダル生成AI

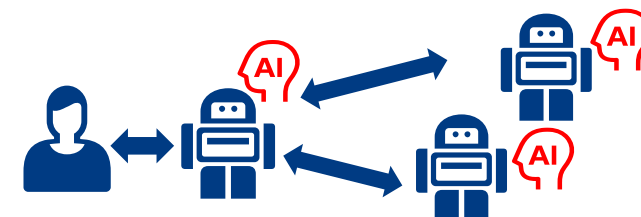
テキスト・画像・音声など複数のモダリティを統合して認識・理解し、生成



活用例：医用画像に基づく説明文生成、各種センサ情報を活用するロボット

AIエージェント

目的の達成に向け、状況に応じて計画を立てつつ、自律的に行動を選択・実行するAI
⇒さらに物理空間での行動制御や運用支援へ

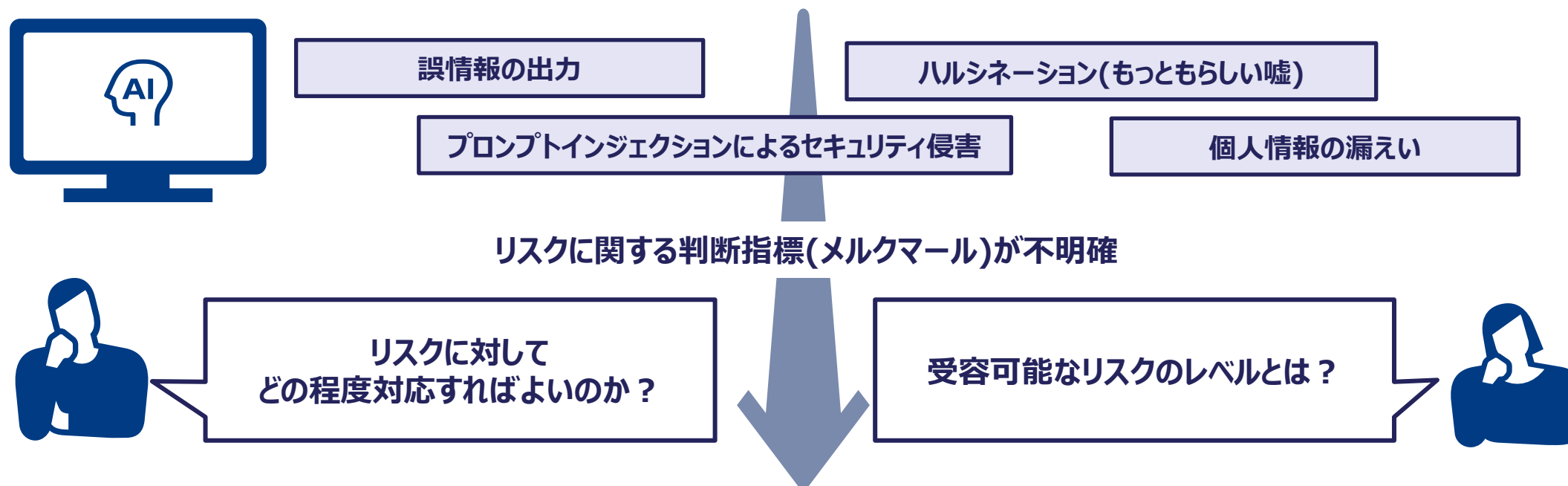


活用例：定型業務プロセスの自動化、意思決定支援・自動対応、災害現場のドローン、工場内の協調搬送ロボット

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。

生成AI利活用によるリスクの顕在化



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

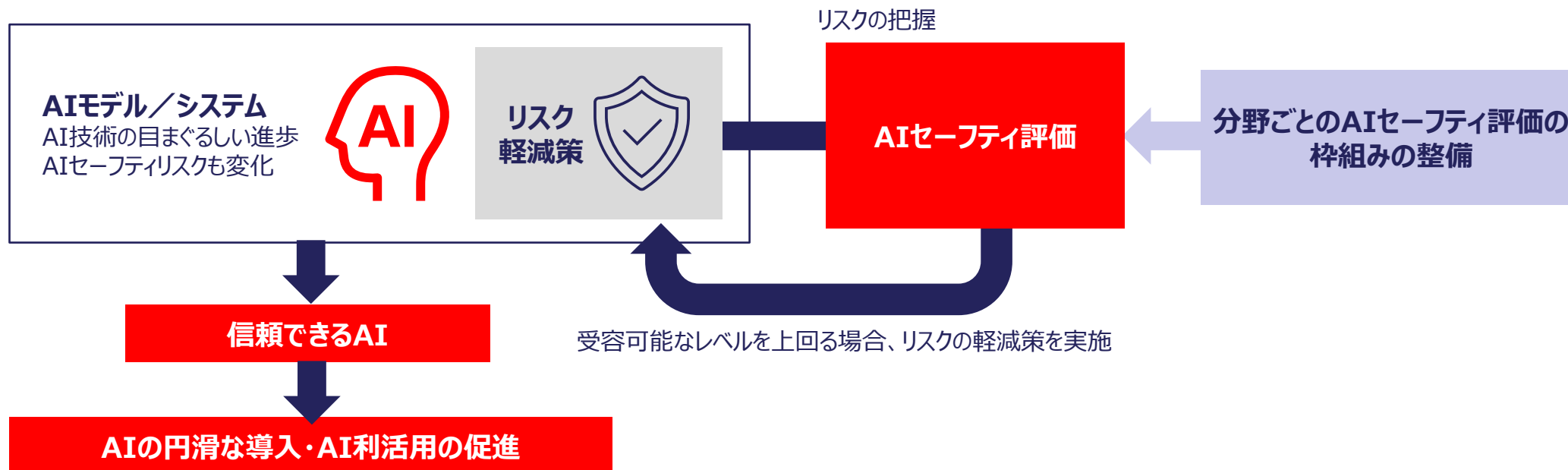
適切なAIセーフティ評価により、AIが社会において、「信頼できるAI」として受け入れられるための基盤を形成

- 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

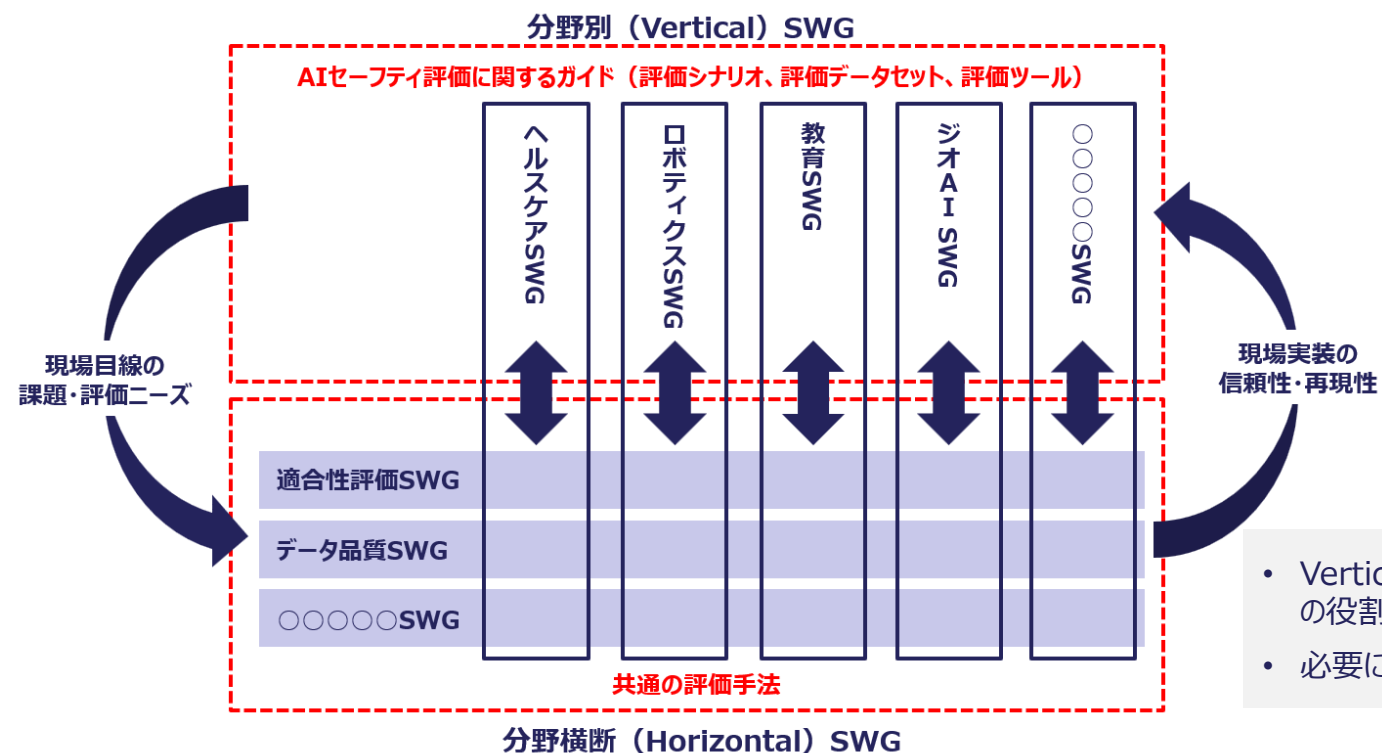
AIセーフティ評価の枠組みの整備を通じて：

- 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。



- **分野別（Vertical）SWG*と分野横断（Horizontal）SWGを組み合わせた体制を構築。**
- 現在、6つのSWG（ヘルスケア、ロボティクス、教育(新)、ジオAI(新)、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



- 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。
- 必要に応じてSWGの増設も計画する。

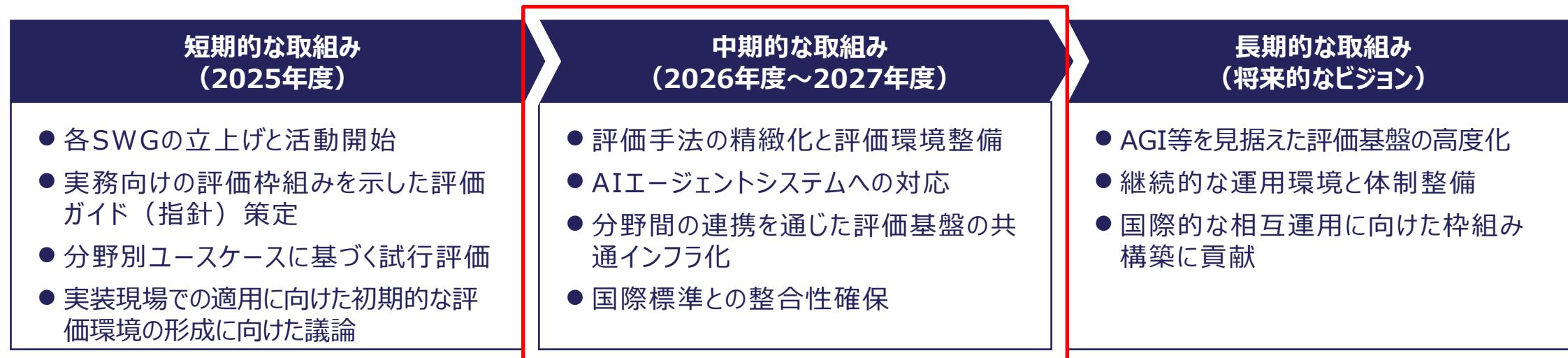
分野別（Vertical）SWG：各分野向けの評価観点ガイドの検討

- 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

- 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。
- 2025年度は、重点分野のユースケースで評価観点の整理や試行評価を実施し、AIセーフティ評価を実務に落とし込む評価ガイド（指針）を策定。
- 2026年度以降、中期的には、現場での適用と運用を通じたフィードバックループの確立と評価手法の精緻化、評価環境の整備等を進め、技術多様性への対応と国際標準との整合を図る。



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

- 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

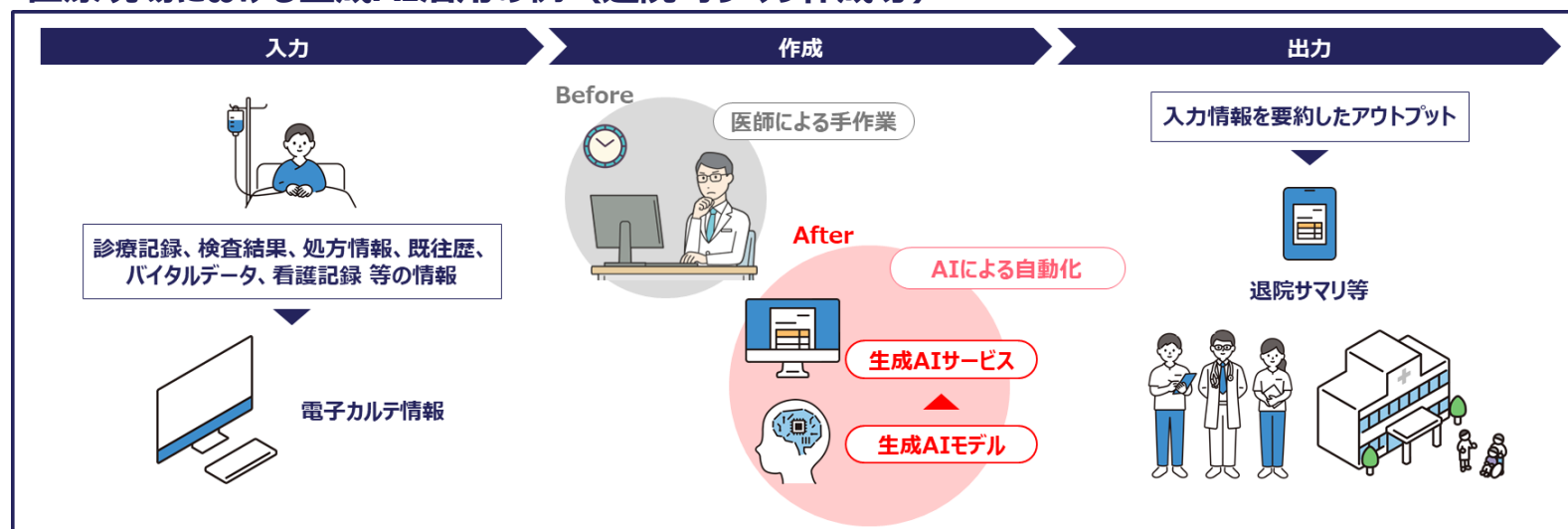
- ・医療機関における時間外労働上限規制
⇒医療従事者の書類作成業務などの負担軽減
- ・一般生活者の健康管理の商品・サービスも普及
- ・分野独自の取り組み：
ヘルスケア事業者のための生成AI活用ガイド*策定

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ガイドにより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
- ・ユースケースごとの評価手法・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）



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。

- 生成AIの出力は患者の生命・身体・精神に直接影響し得るうえ、要配慮個人情報を含む機微なデータを扱うため、ヘルスケア領域固有のリスクへの対応が求められる。
- AIセーフティの10評価観点を適用しリスクを整理した『ヘルスケア領域におけるAIセーフティ評価観点ガイド』を公開。

想定される リスク	● 患者の生命・身体・精神への直接的影響 ⇒ 不正確な出力や不適切な応答がユーザーに直接影響を与える ● 取り扱うデータの機微性 ⇒ 要配慮個人情報やプライバシー性の高い情報等の漏えいや不適切な推論によりユーザーに深刻な影響を与える	AIセーフティ評価 の方向性
		『ヘルスケア領域におけるAIセーフティ評価観点ガイド』(2026年4月公開) ● Non-SaMD*のテキスト生成AIを対象（今後の進展を踏まえ随時見直し） ● AIセーフティに関する10の評価観点についてヘルスケア領域への適用を想定し、リスクを整理 ● プロダクト開発フェーズごとの重要な評価項目をチェックリスト化し事業者のガイド活用を促進

SWGの ロードマップ	短期的な取組み (2025年度)	中期的な取組み (2026年度～2027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 ● 評価観点ガイド第1.0版策定	● AI開発・提供事業者による評価観点ガイドの効果検証 ● AI開発・提供者とAI利用者の観点で具体リスクや課題の整理とガイドへの反映	● 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

- 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 ・「Physical AI」や「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 と利活用促進 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・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。
- 政府のAIロボティクス戦略等を踏まえ、リスク指標の体系化や受容性・信頼性を含む統合的分析手法を確立し、評価モデル・設計指針の共有により、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を目指す。

想定される リスク	・ 物理的リスク ⇒衝突・挟圧・転倒、誤動作 等 ・ 心理的リスク ⇒威圧感・ストレス、依存、 判断の歪み 等 ・ 社会的リスク ⇒権利侵害・不正利用、 不公平・差別的扱い 等 物理的リスク  心理的リスク  社会的リスク 	『AIロボティクスに関するセーフティ評価観点ガイド』（2026年7月公開） ・ AIの利用による新たなリスクをどのように低減するか、具体ユースケースを対象に、開発・提供・運用の各フェーズにおけるリスク整理とシナリオに基づく評価実験を実施し、AIセーフティ評価の観点を取りまとめ ・ 評価実験に基づく、評価観点の見直しと分析手法の確立・高度化を継続 AIセーフティ 評価の方向性
----------------------	---	---

SWGの ロードマップ	短期的な取組み (2025年度)	中期的な取組み (2026年度～2027年度)	長期的な取組み (将来的なビジョン)
	● AIロボティクスにおけるリスク類型を整理 ● 「飲食店等における商品搬送ロボット」および「遠隔操作型の配送ロボット」に関する模擬環境での評価実験 ● 評価実験結果から抽出したAIセーフティ評価観点の取りまとめ	● 「AIロボティクス戦略」および「AIロボティクス実装ロードマップ」を踏まえた実装可能性の高い動作に関する評価実験 ● 評価実験結果にもとづくAIセーフティ評価観点の見直し	● 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

- 教育分野、とりわけ初等中等教育段階では、児童生徒や教師のAI活用の可能性とともにリスクが明らかになってきた
- 安全かつ主体的なAIの利活用の促進に向け、ユースケース毎のリスクを基に、AIセーフティの評価観点の整理が必要

教育分野の AI利活用の 現状

- 教育分野、とりわけ**初等中等教育段階**では、**文部科学省**は、学校現場で生成AIを利活用する際の**考え方や留意事項をガイドラインとして公表**
 - ✓ 『初等中等教育段階における生成AIの利活用に関するガイドライン Ver.2.0』（文部科学省：2024年12月）
- 国内外で、教育分野の様々な場面でAI利活用が進むとともに、児童生徒や教師向けの多様なAIサービスが開発・実装され、**AIの利活用に関して、可能性やリスクが明らかになってきた**
 - ✓ 文部科学省での生成AIパイロット校事業*1や教育分野特化の生成AIの実証事業*2等
 - ✓ OECD『Digital Education Outlook2026』（2026年1月公表） 等

AIセーフティ 評価の重要性 と利活用促進 への期待

- 安心・安全なAI利活用への期待
⇒教育分野でのAI利活用の際は、単に利便性のみならず、学校教育に係る各種法令や教師の教育的意図等、**教育目的に照らして安心・安全に利活用できるかを評価する視点が必要**
- 教育分野のユースケースごとのリスク整理
⇒個別の学習場面等のユースケースのリスクを踏まえ、**AIセーフティの評価観点の整理**が求められる
- 評価基準・評価データ等の整備
⇒AIの安全性とセキュリティのベンチマークの取組や新たな技術等の進展も踏まえつつ、**教育分野における評価観点ガイドや、AI安全性ベンチマーク及び評価環境整備**により、教育分野での安全かつ主体的なAIの利活用の促進につなげていく

*1 文部科学省「リーディングDXスクール事業 生成AIパイロット校」（令和6年度）、「生成AIパイロット校の指定を通じた事例創出事業」（令和7年度、令和8年度）

*2 文部科学省「学びの充実など教育課題の解決に向けた教育分野特化の生成AIの実証研究事業」（令和7年度）、「教育課題の解決に向けた教育分野特化の生成AIの実証事業」（令和8年度）

- 教育分野のリスクの例として、学習過程の過度なショートカット（メタ認知的怠惰）や、依存等が指摘されている
- 目下は、生じ得るリスクの検討を進め、今後は、評価ベンチマーク、AI提供者向けのガイドラインの整備を目指していく

想定 される リスク例

- AIへの情緒的依存
⇒AIを擬人化し、過度な愛着を持つ
- 過度なAI依存によるメタ認知的怠惰
⇒「生産的な試行錯誤」の機会の喪失や、
学習者のエージェンシーの低下を招く
- AI依存による成績低下
⇒学習者が自らの理解・習熟度を過大評価
- プロンプトインジェクションによる評価操作
⇒不正な操作で自身の成績・評価を書き換える

AIセーフティ 評価の方向性

- 教育分野におけるAIプロダクト・サービスを対象に、
発達段階や教育目的、活用場面等を踏まえ、
AIを利活用する際に生じ得るリスクを検討し、
リスクシナリオとして整理・具体化
- ✓ 児童生徒がAIとの対話を通じて説明やヒント、
フィードバック等を得る学習支援や様々な相談
対応の場面等（有識者等へヒアリング実施）
- リスクに対する低減策と、
その有効性を確認する評価観点・項目を整理

SWGの ロードマップ

短期的な取組み (2026年度)

- 教育分野（初等中等教育段階）にお
ける児童生徒の生成AI利活用に係る
ユースケース・リスクの整理
- ユースケース・リスク毎にAIセーフティ評
価観点の取りまとめ

中長期的な取組み (2027年度～、将来的なビジョン)

- 教育分野における児童生徒、学生、教職員を中心とするAI利活用に係るユース
ケース・リスクの整理・更新
- ユースケース・リスク毎にAIセーフティ評価観点の取りまとめ
- 評価シナリオ・ベンチマーク設計（教育分野に特化したデータセットの構築を含む）
- 教育分野のAI提供者向けガイドラインの作成・更新

- 防災、インフラ管理、都市計画等を支える地理空間情報について、AIによる整備・分析・推論への活用を進めていく。
- AIによる分析などによる成果データの品質・信頼性を評価できる枠組みを整備し、ジオAIの適切な導入・利活用を促進していく。

想定される リスク

- 地理空間情報のデータ品質・信頼性
⇒AI活用を前提とした評価手法などが未整備
- AIによる分析・推論結果の信頼性
⇒成果に対する信頼性の枠組みが未整備
- 地理空間基盤モデル
⇒特有の位相関係・空間参照・空間演算などを適切に扱えるAIシステムは発展途上

AIセーフティ評価 の方向性

- 地理空間情報の整備・更新、AIシステム、成果データの評価枠組みを整備
- ユースケースに基づくAIリスク分析と評価フレームワーク・ガイドラインを具体化
- 上記をAIセーフティ評価の具体化取組みとし、国際標準化・ルール形成へ展開、社会環境の変化に伴う新たな課題への対応も検討する

SWGの ロードマップ

短期的な取組み (2026年度)

- 地理空間情報整備・更新用のAIシステム及び成果データの安全性確保、品質評価のための共通フレームワークの検討
- 2次元マップ（建物）を対象とした具体的な指針の検討

中期的な取組み (2027年度～2028年度)

- 2次元マップ（建物）以外の領域への指針検討の拡張検討
- 地理空間情報×AIユースケースにおけるリスク分析、ガイドライン検討
- 国際標準化団体への提案
- 認証の仕組みの検討

長期的な取組み (将来的なビジョン)

- 国際標準として採用
- AIシステム及び成果データの認証の仕組みの運用
- 継続的なルールの見直し、維持

- 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等にて実証する。

現状と課題

- ・ **AIの信頼性確保におけるデータ品質の重要性が増大**
⇒ AIの普及により対象となるデータや利用形態が拡大
- ・ **AI-Ready Dataの整備が課題**
⇒ 意味・来歴等のメタデータも含む品質管理が必要
- ・ **目的・リスクに応じた必要十分で合理的な品質管理**
⇒ 複数の国際標準・枠組みは整備されている一方、具体的な評価方法や導入手順は十分でない
⇒ 目的に応じた品質水準・管理方法が求められる

データ品質の確保に向けた取組み方針

- 『データ品質マネジメントガイドブック』のアップデートと評価ツールの提供
 - ✓ 評価観点と実践フレームワークを整理
 - ✓ 事業者が活用できる評価ツールを提供し、実践上の障壁を低減
- 「評価尺度（メジャー）」の整備
 - ✓ データ品質を共通の基準で評価・比較
 - ✓ ユースケースやリスク特性を踏まえ、現場での適用検証を通じて実効性を確認

SWGのロードマップ

短期的な取組み (2025年度)

- ユースケースの特性に応じて必要なデータ品質の取組整理
- データ品質マネジメントチェックリストの作成・検証
- データ品質の違いが生成AIの応答と安全性に与える影響検証（官報データを題材として活用）

中期的な取組み (2026年度～2027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

- 分野別SWGの取組みを通じて、自己適合評価を含むAIに関する適合性評価手法の確立を目指す。
- 国際標準化の議論を踏まえて、包括的なアシュアランス・アカウントビリティ確保に係る適合性評価制度を設計する。

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は適合性評価のハイレベル枠組み（国際標準化が進行中）
⇒柔軟かつ適切な評価方法と実行可能な手順が必要

適合性評価手法の確立への取組み方針

- 他SWGの活動から、AIへの要求事項等を導出
- 自己適合評価を含むAIに関する適合性評価手法と実行可能な運用手順の確立と制度設計を目指す

<具体的なステップ>

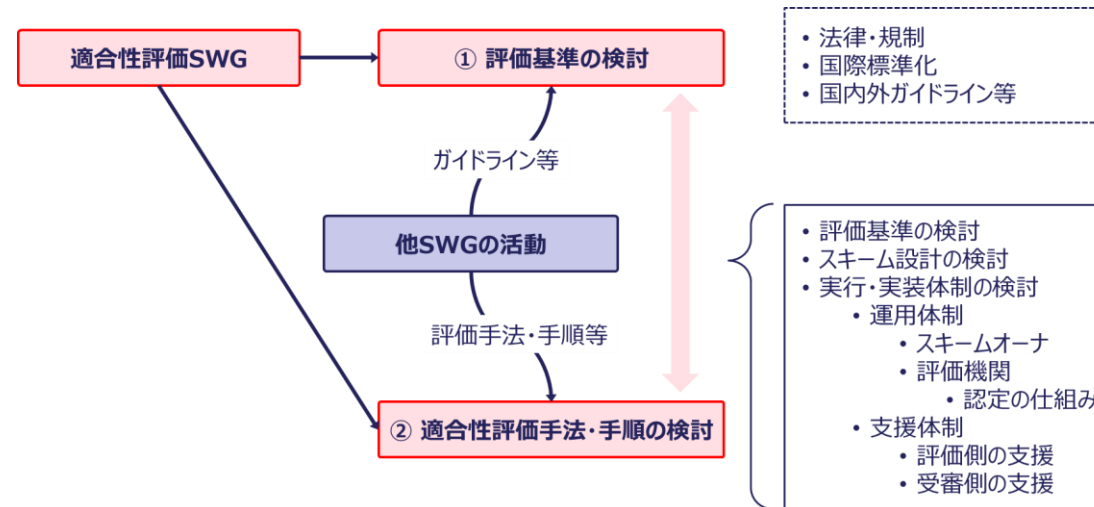
ヒアリング・課題抽出

制度設計に向けた検討

ツール・体制整備

国際標準との整合

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 (2025年度)

- 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するとともに、包括的なAI適合性評価について検討
- 分野別SWGとの連携によるヒアリング

中期的な取組み (2026年度～2027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

AISI

Japan AI Safety Institute