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～

令和 7 年 7 月 10 日

AISI Japan AI Safety Institute

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

目次

0. 本書の位置付け

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機能することを目標としており、AI セーフティ・インスティテュート（AI Safety Institute: AISI）の実証活動と連動しながら、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ものである。

本ビジョンペーパーにおける「AI セーフティ」の定義は、2024 年 9 月に AISI が公表した『AI セーフティに関する評価観点ガイド』に記載されている通り、「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）」である。

AI セーフティを適切に評価することは、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し、信頼とイノベーションの両立を実現する上で不可欠な要素となる。

本ビジョンペーパーが、AI の安全かつ有効な利活用を目指す事業者、研究者、政策担当者等が、例えば AI セーフティに関する評価の枠組みや評価項目の設計を行う場合の参考として、広く参照されることを期待する。特に、AI 開発・提供事業者（2024 年に総務省・経済産業省が策定した『AI 事業者ガイドライン』[1]における「AI 開発者」および「AI 提供者」をいう。）においては、事業者自身が開発または利用する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面を想定している。また、AI の導入・利活用を行うユーザー企業や自治体等においては、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されることが期待される。さらに、政策担当者や認証機関等においては、今後の評価指針の策定等に向けた議論に資することが期待される。このように、評価に関与する多様なステークホルダーがそれぞれの立場で参照することで、AI セーフティ評価に関する共通理解の醸成と社会実装の促進に貢献することを目的としている。

1. はじめに

近年、生成 AI をはじめとする AI 技術の急速な進展は、産業・行政・医療・教育など多様な分野において、従来の業務プロセスなどに大きな変革をもたらしている。とりわけ、大規模言語モデル（Large Language Models: LLM）に代表される自然言語処理技術は、人間の意思決定や創造的なタスクへの応用可能性を飛躍的に高め、社会的影響と経済的インパクトの両面において前例のない注目を集めている。さらに、テキストのみでなく、画像、音声など様々なモダリティのデータを一元的に扱うことのできる AI モデルは、多様な AI アプリケーションの共通基盤として機能する。今後、社会のデジタルインフラやサービスの多くがこうしたモデルに依存する構造になり、その上に個別の AI サービスやエージェントが構築されていく可能性がある。我が国では、少子高齢化による労働力不足が喫緊の課題となる中、AI は有力な対応手段として期待されており、その利活用の促進は社会・産業の持続的成長に不可欠とされている。

一方で、AI が人間の行動や判断、さらには公共的な意思決定にまで影響を及ぼすようになる中で、誤情報や偏見の生成、ブラックボックス性、説明困難性といった新たなリスクも顕在化している。これらのリスクは、ビジネスに重大な影響を与える可能性があるため、導入に慎重な姿勢がとられ、AI 利活用が十分に進んでいない現状がある。「利便性と効率性の向上」と「不確実性と社会的リスクの増大」が同時進行する状況において、AI セーフティを確保しながら利活用を進めることは、我が国を含む国際社会に共通する重要課題である。この課題の解決には、AI を利活用したイノベーションの推進と並行して、適切なリスクマネジメントの実施が必須となる。

こうした状況を踏まえ、日本政府は AI のリスクに適切に対応しつつイノベーションを加速するため、「AI セーフティ」に関する評価手法や基準の検討・推進を担う機関として、2024 年 2 月に AI セーフティ・インスティテュート（AI Safety Institute: AISI）を設立し、同年 9 月にはその基本的な考え方を示す各種ガイドを策定・公表した。さらに、2025 年 3 月にその普及と具体化に向けて「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）」を新設し、G7 広島 AI プロセスでの国際的合意、『AI 事業者ガイドライン』（総務省・経済産業省）、AI に関する国際標準動向（ISO/IEC JTC1 SC42 等）を踏まえつつ、実証等を通じて具体的なユースケースに即した評価ガイド（評価シナリオ、評価データセット、評価ツールを含む）の体系化に着手している。事業実証 WG では、民間事業者を中心に多様な主体が参画し、実務的な視点からユースケースやニーズを発掘し、民間の知見や課題を議論し体系化する。民間事業者の機動力を活かし、業界や利用者の立場に根差した評価ガイド策定を通じて、実践的なルール形成が進むことが期待される。

事業実証 WG の活動は、AI の社会実装と産業展開に資する AI セーフティ評価の確立を目指すものであり、政府関係機関のみならず、民間企業や技術者の取組みを支えることを狙いとしている。また、分野ごとに異なるリスク構造や評価ニーズを踏まえつつも、分野横断的な評価手法の整備やデータ品質・適合性評価の整備により、日本発の実践的で重層的な評価枠組みの構築を目指し、国際的・標準的なルール形成にも寄与することを視野に入れている。

2. 事業実証 WG のビジョン

AI の社会実装を加速するためには、実際のビジネス・ユースケースを前提としたリスク評価と実証の両輪による検証が不可欠となっており、事業実証 WG ではそのための活動の場を提供していくことを想定している。特に、高度な判断や出力を伴う生成 AI の活用が進む中、分野別・機能別に応じた実用的なユースケースを通じて、評価手法を確立することが求められている。

本章では、こうした背景を踏まえ、「信頼に足る AI」の利活用を支える AI セーフティの評価の枠組みの構築に向けた事業実証 WG の活動の方向性を示す。

2.1. AI 利活用の現状

近年の AI 技術の進化は著しく、とりわけ LLM と生成 AI の出現は、従来の AI 利活用のあり方を根底から変えつつある。2022 年末に登場した ChatGPT は、日常会話から業務支援、創作活動に至るまで多様な用途に用いられるようになり、一般市民から産業界まで幅広く AI の恩恵を受けられる環境が整いつつある。生成 AI は、自然言語生成や自動要約、翻訳、分類などを通じて、多くの業務プロセスの自動化や支援に貢献している。近年では、これらの基盤的な自然言語処理能力を活用し、実務における高度な情報処理や判断支援を実現する応用技術が注目されている。(図 2-1)

生成 AI の重要な応用技術の一つとして、外部知識拡張（External Knowledge Augmentation）の代表的手法である検索拡張生成（Retrieval-Augmented Generation: RAG）が挙げられる。これは、AI がリアルタイムで外部データベースなどから必要な情報を検索し、その結果に基づいて出力を生成する仕組みである。特に、企業内部のナレッジを活用した FAQ 応答や業務支援ツールへの導入が進んでおり、ドメイン特化型 AI の発展にも寄与している。次に、マルチモーダル生成 AI は、テキスト・画像・音声など複数のモダリティを統合して処理・生成するものであり、より柔軟な情報理解と表現を可能にする。異なる情報ソースを組み合わせることで、文脈把握や状況判断の精度が向上し、人間に近い認知的処理が可能となる。例えば、医用画像に基づく説明文の生成、各種センサの情報を組み合わせて人間の五感に基づく認識や判断を代替するロボットなど、実環境に即した活用が期待されている。さらに、自律的にタスクを遂行し、状況に応じて判断・行動する AI エージェントの進展も著しい。AI エージェントは、予め定められた目的や業務フローに従って処理を行う「ワークフロー型」と、外部環境の変化を踏まえて自律的に判断・行動する「自律型」に大別される。ワークフロー型 AI エージェントは、定型的な業務プロセスの自動化ツールとして既に広く活用が進んでおり、業務の効率化や標準化に寄与している。一方、自律型 AI エージェントは複雑な情報環境下における意思決定支援や自動対応といった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されている。例えば、災害現場で探索を担うドローンや、工場内で複数のロボットが協調して搬送を行うシステムなど、様々な現場での実装が現実味を帯びつつある。(図 2-2)

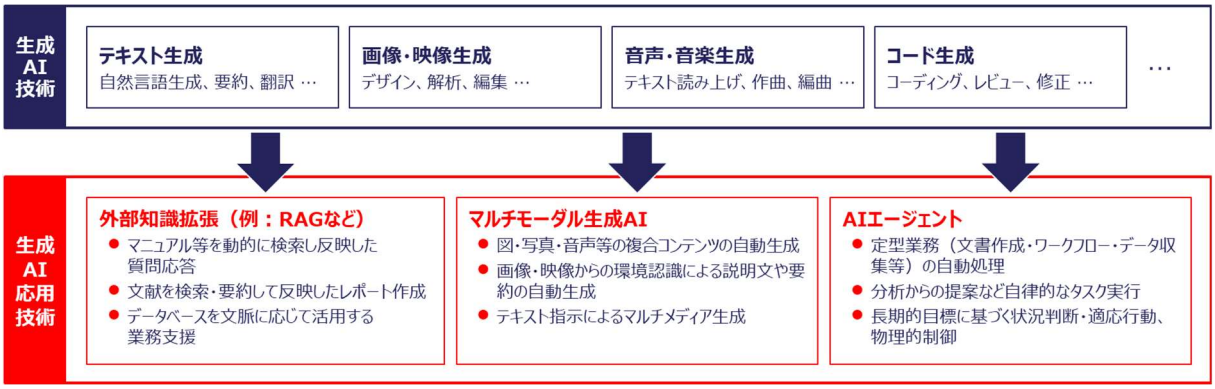

図 2-1 AI の技術分類と用途

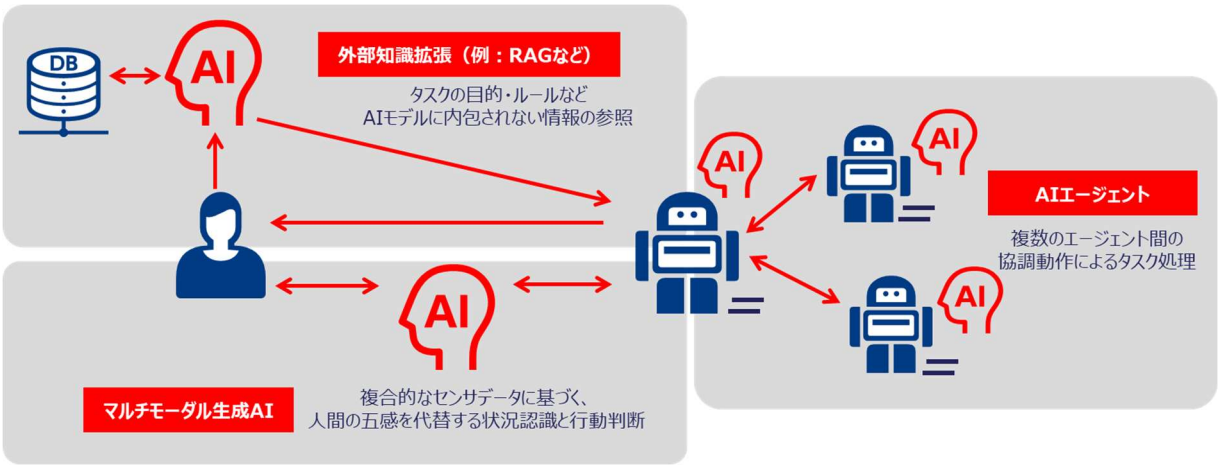

図 2-2 物理空間における生成 AI 応用技術の利活用イメージ

このように、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も質的な変化をもたらしている。

従来、AI は限定的な判断や処理を担う補助的な要素として導入されてきたが、現在では、情報の取得・統合・解釈を含むプロセス全体を担い、他のシステムや人と連携しながら意思決定を行う構成要素として実装される場面が増えている。これにより、AI は単体で機能するモジュールから、複数の機能と連動し統合的に判断・行動する存在へと変化しつつある。実際に、AI の利活用は文書作成業務や対話業務にとどまらず、医療（例：退院時サマリ生成や患者説明支援）、製造（例：異常検知 AI と連携した巡回点検ロボットによる予防保全）、流通・接客（例：施設内を移動しながら対話や案内を行うサービスロボット）など実世界の応用領域へと急速に拡がっている。

また、これまで人の直感や経験に依存してきた領域においても、AI による補完的支援の可能性が高まりつつある。曖昧な判断、感覚的理解、臨機応変な対応といった、従来は機械化が困難とされていた要素にも対応可能な技術が登場し、現場レベルでの活用が現実のものとなりつつある。これらの動きは、AI が社会インフラとして機能する新たな段階に入っていることを示している。

こうした AI の技術進展や役割の変化等に呼応する形で、『AI 事業者ガイドライン』をはじめ、産業界・国際社会においても AI に関するルール形成が進んでいる。同時に、イノベーションの促進と「信頼に足る AI」の利活用の両立を目指し、リスクの洗い出しや責任の所在明確化などの検討が本格化しており、AI セーフティ評価のあり方を具体的に考える基盤が整いつつある。

2.2. AI セーフティ評価の必要性

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している。特に、生成 AI による誤情報の出力やプロンプトインジェクションによるセキュリティ侵害、個人情報の漏洩といった問題が現実に発生しており、生成 AI の出力に起因して企業の評価が毀損した事例や、生成 AI を用いたチャットボットの利用が原因で人の身体・生命に影響が生じた海外事例も報告されている。

さらに、生成 AI の出力はしばしば「もっともらしい嘘（ハルシネーション）」を含む。これは、モデルの性質や学習方法、学習データの品質など複合的な要素に起因するものであり、完全に防ぐことは困難である。

こうしたリスクに対しては、AI セーフティ評価によってリスクの特性と影響を適切に把握することが重要であり、その上で、RAG のような外部情報検索の活用や、アウトプットの根拠になる引用元明示など、技術的な対応をはじめとする機能的リスクマネジメントによって対応していくことが現実的な手段である。

一方で、安全に AI を利活用するためにどの程度の対応が必要かという判断指針（メルクマール）が明確でないため、AI の導入をためらう現場も少なくない。リスクの洗い出しや対策は進みつつあるものの、それをどのように評価し、どの水準で「社会的に受容可能」とみなすかという枠組みが整っていないことが、実装における大きな障壁となっている。

AI が社会に広く受容されるためには、これらのリスクをゼロにするのではなく、社会的に受容可能な水準に制御することが求められる。そこで、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ことが急務となっている。

2.3. 政策対応と分野別の取組み

2.3.1. 関連する政策・規制動向

現在、AI セーフティに関する政策・制度整備は世界的に加速しており、国や地域ごとに制度設計の方針や規制の厳格度は異なるものの、共通して「リスクベース評価」「透明性確保」「責任分担の明確化」などを重視する傾向が見られる。（表 2-1）

我が国においては、2024 年に『AI 事業者ガイドライン』が策定され、特に民間事業者主体による共助的なガバナンスの枠組みと、実装支援を重視する政策的スタンスが打ち出された。これを受け、AISI は、同ガイドラインをベースに、『AI セーフティに関する評価観点ガイド』[2]と、評価手法の一つである『AI セーフティに関するレッドチーミング手法ガイド』[3]を公開した。レッドチーミングガイドの別添では、エンタープライズ向けの RAG を利用した社内業務データ活用チャットボットサービスに対する脆弱性の発見と、その改善策が示されている。

さらに、2025 年 6 月には、AI 推進法（人工知能関連技術の研究開発及び活用の推進に関する法律）が公布され、AI の研究開発・活用の推進に向けて、事業者を含む国全体の責務と基本的施策が定められた。これらに呼応する形で、AISI では分野別サブワーキンググループ（SWG）によ

るユースケースに基づくリスクの洗い出しや、AIセーフティ評価手法の具体化に取り組んでいく。なお、ヘルスケアやロボティクスなど分野別 SWG の対象となる分野には、既に関連する法律や業界ガイドラインが存在するケースが多く、AI セーフティ評価の設計にあたっては、これらの既存の制度やルール体系との整合性・補完性を適切に考慮する必要がある。

一方、国際的な動向としては、EU が AI Act の制定を先行して進めており、リスクベースに応じた義務の層別化（例：高リスク AI に対する第三者適合性評価（認証）義務など）を制度化している。米国はこれと対照的に、米国国立標準技術研究所（National Institute of Standards and Technology: NIST）の AI リスクマネジメントフレームワーク（AI RMF）で、法制化よりも AI を通じた自発的・文書化を重視する非規制的アプローチを採用しており、官民連携による柔軟な実務展開が進められている。

また、AI マネジメントシステム（ISO/IEC 42001）やインパクト評価（ISO/IEC 42005）といった国際標準が既に発行され、適合性評価のハイレベル・フレームワーク（ISO/IEC 42007）などの新たな国際標準の策定も進行中であり、日本もそれらの議論に積極的に参画している。

プロセスに着目する柔軟性や、リスクカテゴリに基づく厳格さの両面がある中、日本の AI セーフティ評価に資する環境整備に向けては、より多くの産業や企業に実装可能であり、かつ国際標準との相互運用性を確保することが、今後の重要な検討課題となる。各国の制度動向を踏まえた柔軟かつ実効的な評価の枠組みの構築が、日本の AI セーフティ戦略における鍵を握ると考えられる。

表 2-1 主要国の AI 制度と評価アプローチ比較

	日本	欧州	米国	英国	中国
規制手法アプローチ	**ソフトロー リスクベース**	**ハードロー リスクベース**	**ソフトロー 自発的・文書化**	**ソフトロー 原則・規範ベース**	**ハードロー 管理モデルベース**
原則	人間中心のAI社会原則 （2019年）	AI・ロボティクス・自律システムに関する宣言 （2018年）	AI権利章典 （2022年）	国家AI戦略 （2021年） AI白書（AI規制におけるイノベーション促進型アプローチ） （2023年）	次世代人工知能倫理規範 （2021年）
指針・ガイドライン	AI事業者ガイドライン （2024年）	信頼できるAIのための倫理ガイドライン （2019年）	AIリスクマネージメントフレームワーク（AI RMF） （2023年）	英国のAI規制原則の実施規制当局向け初期ガイダンス （2024年） 英国政府のための人工知能プレイブック （2025年）	信頼できる人工知能についての白書 （2021年） 人工知能倫理ガバナンス標準化ガイドライン （2023年）
政策など	AI推進法 （2025年）	AI法（AI Act） （2024年）	人工知能の安全・安心・信頼できる開発と利用に関する大統領令 （2023年） ※ただし、2025年に撤回	AI（規制）法案 （2025年上院提出）	生成人工知能サービス管理暫行弁法 （2023年）

2.3.2. 分野ごとの取組み

AI 利活用にあたっては、政府主導のルール形成だけでなく、民間業界団体等による自主的な取組も進展しており、『AI 事業者ガイドライン』における 3 本柱の一つである「自主的取組の推進」の具体化が進められている。

例えば、ヘルスケア分野では、日本デジタルヘルス・アライアンス（JaDHA）が『ヘルスケア事業者のための生成 AI 活用ガイド（ヘルスケア領域において生成 AI を活用したサービスを提供する事業者が参照するための自主ガイドライン）』（2025 年 2 月第 2.0 版公表）[4]を策定し、ヘルスケア事業者が生成 AI を活用してサービス・プロダクトを提供する場面を想定した実務的な手引きとして提示している。また、一般社団法人 AI メンタルヘルスケア協会（AIMH）が中心となりメンタルヘルスケアのようなセンシティブな領域における生成 AI の活用に関して、倫理的・技術的観点からの評価のあり方について、専門家と事業者が連携してガイドラインの策定を進めている。2025 年 4 月には、『メンタルヘルスケア事業者のための生成 AI 活用ガイドライン（案）』が公表され、プライバシー保護や心理的影響への配慮を含めた原則が提示された。

ロボティクス分野では、ユーザーの心理的反応や行動特性、社会的関係性の形成プロセス等を研究する Human-Agent Interaction（HAI）や Human-Robot Interaction（HRI）分野においてサービスロボットの対人インタラクションに関する評価の枠組みが整理・蓄積される中、AI の進化により評価対象は広がりつつある。また、環境の認識から動作の生成までを一つのモデルで扱える End- to -End モデルによる開発スタイルの登場により、ロボット設計そのものにも根本的な変化の兆しが見られる。

さらに、金融分野では、一般社団法人金融データ活用推進協会（FDUA）が『FDUA 生成 AI ガイドライン』（2024 年 10 月第 1.0 版公表）[5]を策定し、生成 AI の適正利用に関するリスク評価の枠組みや、法令遵守・説明責任の確保に資する実践的な留意事項を提示している。加えて、金融機関各社においても、内部ガバナンス体制の強化や社内ルールの整備が進められており、生成 AI の利活用と信頼性確保の両立を図る動きが見られる。

このほか、教育、製造、自治体分野などにおいても、各現場の特性に応じた AI 利活用に向けたガイドラインの策定や、評価観点の明確化が進められており、業種・業態を問わず、民間主導による AI セーフティの確保にかかわる取組みが着実に広がりを見せている。

これらの民間主導の取組みは、AISI において整備された『AI セーフティに関する評価観点ガイド』等と連動する形で、AI セーフティ評価のベースラインを踏まえた分野別・文脈別のチューニングの取組みと位置付けられる。事業実証 WG では、こうした既存の民間ガイドラインの知見も踏まえつつ、分野別 SWG において実証と検証を重ね、各分野における実装可能な評価基準の高度化を目指している。

2.4. 本事業の目的と実施体制

2.4.1. 目的・ゴール

事業実証 WG 活動の根底にある目的は、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することである。特に、単なるチェックリスト型の対応にとどまらず、利用者のリスク理解を前提とした説明可能性・適合性・データ品質などの多面的評価を可能とする枠組みを整備することが本 WG の核心的ゴールである。

さらに本 WG では、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 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こと、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こと、を中長期的な目的として掲げている。

こうした評価の枠組みは、「信頼に足る AI」のモデル要件を明確にし、AI 開発・提供事業者が自らそれを担保し、利用者や規制当局に対して透明性のある形で説明・証明できるようにするためのものである。また、評価結果が過剰な負担や形式主義に陥ることなく、現場での導入促進や品質向上のインセンティブとして機能するよう、段階的かつ柔軟な構成が求められる。

事業実証 WG では、『AI セーフティに関する評価観点ガイド』等を踏まえ、こうした枠組みの構築を進めるとともに、産業分野ごとの評価項目の抽出や、分野横断領域（例：データ品質・適合性評価）との連携により、AI の社会実装につながる AI セーフティ評価のモデル提示を目指す。

本体系により、リスクを抑える「守りのセーフティ」にとどまらず、日本が強みを持つヘルスケアやロボティクス等の分野、および QCD（品質、コスト、納期）での優位性を活かし、AI セーフティに関する合理的な説明責任を果たす評価手法の枠組みにおいて、国際的に先導的役割を果たすことを目指している。それらの枠組みを、AI によるイノベーションを促進するための「攻めのセーフティ」として積極的に活用することで、国際競争力の向上につながる戦略的なアプローチを推進する。「攻めのセーフティ」とは、受容可能なリスクに抑えるための合理的な考え方や手法を明確化し、それを社会的な説明責任を果たすための手段として用いることで、過剰な安全対策や不確実性によるコストを抑制しつつ、「信頼に足る AI」の利活用と実装を加速させる枠組みである。（図 2-3）

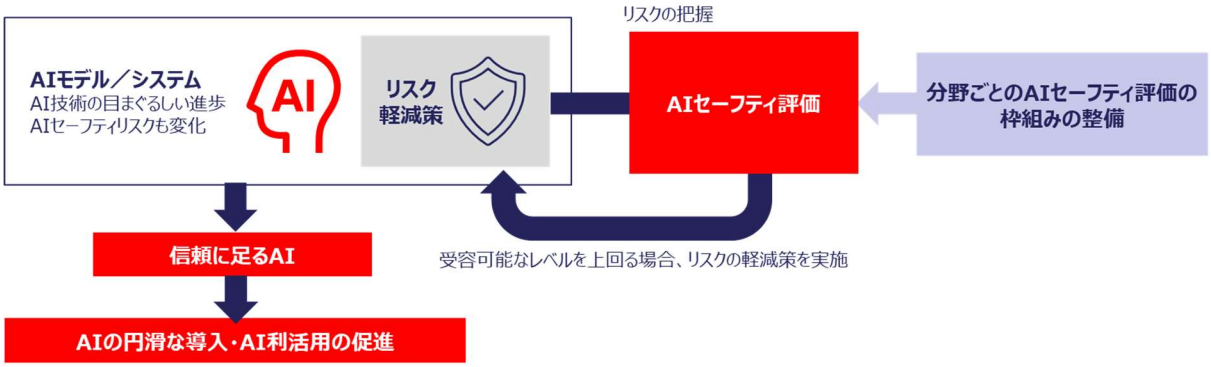

図 2-3 事業実証 WG の活動の全体像

2.4.2. 事業実証 WG の実施体制

事業実証 WG では、「信頼に足る AI」を実現する実効的な AI セーフティ評価の枠組み構築に向けて、複数の SWG を設置している。具体的には、各分野向けの評価ガイド（評価シナリオ、評価データセット、評価ツールを含む）の検討を通じて実装に取り組む分野別（Vertical）SWG と、分野共通の評価手法を検討する分野横断(Horizontal)SWG を組み合わせた体制を構築している。

分野別 SWG は、国際的にも関心が高く、特に日本が技術的・産業的に強みを有する分野のうち、AI 利活用の進展とともに AI セーフティの確保が特に重要となる分野として、ヘルスケア分野およびロボティクス分野を対象に、AI セーフティ評価に関する事業実証を行う。

- ヘルスケア SWG：
  ヘルスケア分野は、医療機関における業務支援ツールや一般生活者向けのチャットボットサービスをはじめとした、多様なサービスが創出され、各々の領域で課題解決に貢献する事例が創出されるなど AI 利活用が進展している分野である。一方で、プライバシー性の高い情報を取り扱う場面が他分野よりも多いため、ヘルスケア分野における AI セーフティ評価項目や方法を明確化することで、安心・安全な利用環境の整備と AI の社会実装を後押しすることが期待されている。
- ロボティクス SWG：
  ロボティクス分野は、労働力不足への対応や現場業務の高度化といった社会的ニーズを背景に、AI の導入に対する期待が高い。特に、視覚・言語・動作を統合する視覚言語行動モデル（Vision Language Action Model: VLA）などの進展により、自然言語による指示に基づく多様な環境での物理的タスクの自動実行が可能になりつつあり、対人インタラクションを伴うシーンにおける安全性の確保が重要となっている。そのため、ロボティクスの機能安全の枠組みを踏まえた AI セーフティ評価の枠組みを明確にすることで、さらなる利活用の広がりが見込まれる。
- データ品質 SWG：
  現在の AI は、大量かつ多様なデータを用いて学習するデータ・ドリブン型であり、その性能や信頼性はデータの質に大きく依存している。こうした背景から、AI の品質を支える基盤として、データ品質を確保するための評価観点や実践的なアプローチを明示することが求められている。
- 適合性評価 SWG：
  AI に関するリスクマネジメントや説明責任の枠組みが国際的に整備されつつある中で、それらに準拠しているかを確認する実効性のある評価手法に対する関心が高まっている。こうした要請に応え、柔軟かつ現場に適した適合性評価のアプローチを提示することが期待されている。

各 SWG は、WG 運営事務局の支援のもと、評価観点の整理、ガイドラインやツールの整備、他 SWG との連携・整合を行い、今後の活動本格化に向けた基礎的素材として蓄積していく。また、必要に応じて SWG の増設を計画する。

分野別（Vertical）SWG と分野横断（Horizontal）SWG は、それぞれの役割を担いつつ、相互に補完・連携しながら AI セーフティ評価の枠組みを形成していく。ヘルスケアやロボティクスといった分野別 SWG では、実際のユースケースを起点に、現場で直面する具体的な課題や評価ニーズが抽出され、それを一般化することで、分野横断 SWG における評価項目やツール整備の方向性が定まる。一方、分野横断 SWG で構築された評価指針は各分野に適用され、現場実装の支援となる標準的評価指針として機能することを目的とする。こうした双方向の連携により、評価の枠組みは現場性と汎用性の双方を備えたものへと発展し、AI セーフティ評価を社会に根付かせるための実装可能な基盤となる。さらに、実証を通じて得られた知見やモデルケースを提示することで、我が国独自の基準や手法を国際的な枠組み等に貢献していくことを目指す。（図 2-4）

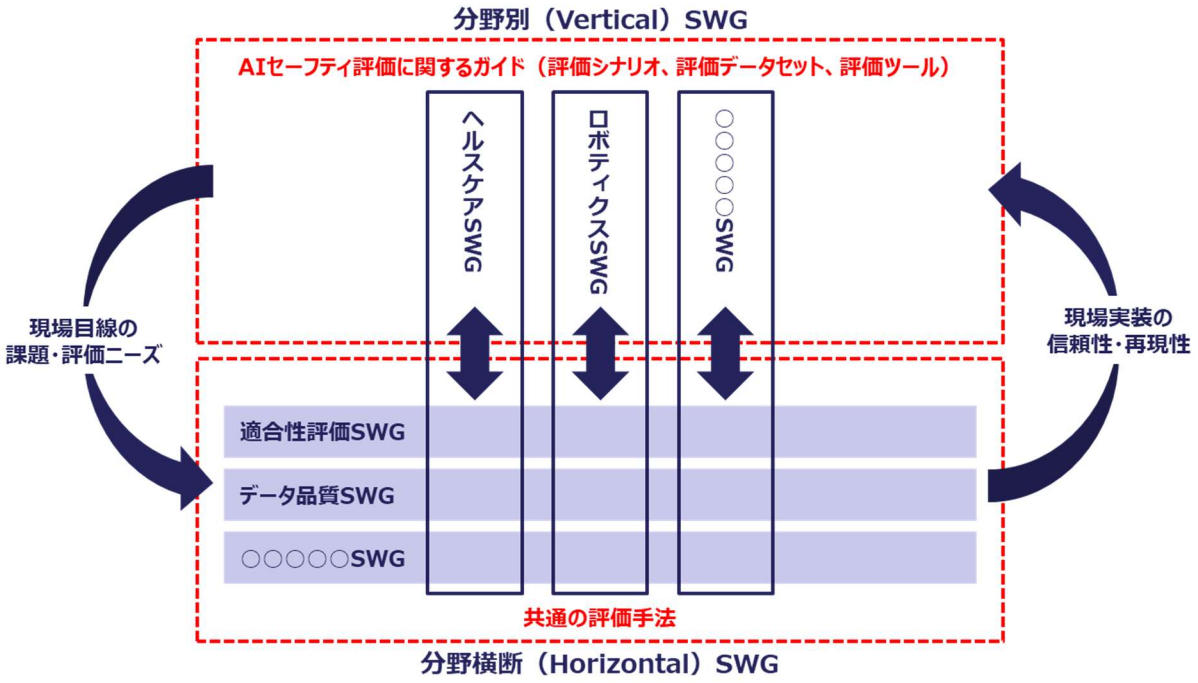

図 2-4 分野別（Vertical）/分野横断（Horizontal）の俯瞰構造図

2.4.3. ロードマップ

事業実証 WG では、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、段階的かつ継続的に取組みを進めていく。

令和 7 年度は、各 SWG の活動を通じて、定性・定量の両面から評価可能な枠組みを提示する。ヘルスケアやロボティクスといった重点分野における具体的なユースケースを踏まえた評価観点の整理や試行評価を実施し、分野横断的に共通するリスク要素も抽出しながら、実装現場における評価の適用に向けた初期的な評価環境の形成を目指す。

中期的には、評価ツール・データセット等を基盤として、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、評価環境の共用・検証機能の整備等を進める。とりわけ、マルチモーダル生成 AI への対応をはじめ、AI エージェントの発展に伴い生み出される新たな技術など、評価対象の技術的多様性にも対応していく。また、社会的影響の大きい分野等への横展開を進め、SWG 活動の適用領域を段階的に拡張していく。分野間での共通課題や評価観点

の調整を図りながら、関係者間での継続的な連携と相互活用を促進し、評価ツールやデータ基盤を共通インフラとして発展させる。これにより、多様な現場への展開を可能にしつつ、国際標準との整合性確保と評価手法に関する国内外の共通理解の形成を目指す。

長期的には、AGI（Artificial General Intelligence）等の実現に付随して生じる新たな AI セーフティリスクも視野に入れた評価の継続的な運用に向け、関連制度との連携を見据えた知見の蓄積と基盤の高度化を進める。様々な分野・領域での本格的な導入を見据え、継続的な評価・監査・運用を一体的に実現する運用環境や、非連続的に変化する技術やリスクに柔軟に対応可能な評価体制を整備していく。さらに、日本発の評価の枠組みを基盤として、国際的な相互運用の枠組みや運用ルールの構築に貢献し、AIセーフティ評価に関する国際的な信頼の確立に寄与することで、我が国のリーダーシップを確立していく。（図 2-5）

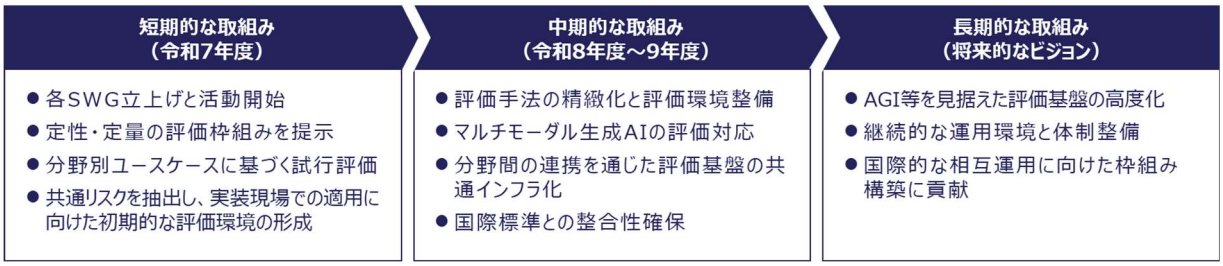

図 2-5 事業実証 WG のロードマップ

3. SWG のビジョン

本章では、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の構築に向けて、各 SWG（分野別・横断的）で検討されている取組みの全体像を整理する。事業実証における AI セーフティ評価は、AI モデルだけでなく AI システムを対象とした評価が必要になるが、このレイヤーを含めた評価の枠組みはまだ確立されていない。このため、分野別・具体的ユースケースごとに、AISI の各種ガイド等を踏まえ、AI セーフティについて検討するアプローチをとることで、評価の納得性を高め、安心して使える AI モデル／システムの評価の実現につなげていく。特に、評価シナリオ、評価データセット、評価ツールを含む分野別の AI セーフティ評価に関するガイドの整備においては、ヘルスケア・ロボティクス・データ品質・適合性評価といった各領域の進展を踏まえ、相互補完的に機能する活動のあり方を描き、自己適合性評価を含む AI セーフティの適合性に関する分野別コンセンサスを確立できるような枠組みを提供する。

3.1. ヘルスケア SWG

2.4.2 に示した通り、ヘルスケア分野は、医療機関向けの業務効率化ツールや一般生活者向けのチャットボットなど多様なサービスが生まれ、課題解決に貢献する事例も登場するなど、AI 利活用が進んでいる領域である。一方で、他分野と比べてプライバシー性の高い情報を取り扱う場面が多いため、AI セーフティに関する評価の観点や評価方法を明確化することが求められる。これにより、安心・安全な利用環境の整備と、AI の社会実装を後押しすることが期待されている。

本節では、まずヘルスケア分野における AI 利活用の現状を概観しつつ、具体的なユースケースを想定した AI セーフティ評価に係る検討課題や取組み指針を取りまとめる。

3.1.1. AI 利活用の現状

近年、生成 AI による自然言語処理技術の進展を中心とした AI 利活用は、ヘルスケア分野にも急速に波及しつつある。

特に我が国では、高齢化社会、医療需要の逼迫等の課題に直面する中、2024 年 4 月から施行された働き方改革法施行による医療機関における時間外労働の条件規制等を背景として、医療従事者の業務効率化や生産性向上に貢献するアプローチの一つとして医療現場でのデジタル技術の導入・活用をはじめとする医療 DX 推進が期待されている。このような状況の中、特に生成 AI は、医療従事者の書類作成業務など従来の業務負担の軽減を実現し、患者の診察やコミュニケーションといった医療従事者の本来業務への専念をより一層に可能にするものとして、大きな注目を集めている。実際、国内の医療現場では、退院時サマリの作成など、医療従事者の手作業で行われていた書類作成業務が大きな負担となっており、医師の時間外労働の理由として「事務作業（記録・報告書作成や書類の整理等）」が「患者対応、ケア」に続いて第 2 位（59.8%）となっている

[1]。

また、ヘルスケア分野においては、継続的かつ個別化されたセルフケア・セルフメディケーションへのニーズの高まりや、PHR（Personal Health Record）データの蓄積が可能なプロダクト・サービスの普及（ウェアラブルデバイス等）、さらに、個人の健康意識・ウェルビーイング志向の高まり等を背景として、生成 AI を用いた個人が自分自身の健康管理を行うための商品・サービス（健康管理アプリ等）の展開等も進んでいる。こうした一般生活者向けの領域は、今後さらに低廉化・利便性の向上が進み、生活の中に普及していくことが見込まれる。ヘルスケアにおける AI 市場規模は、2025 年の 392 億 5,000 万米ドルから、2032 年までに 5,041 億 7,000 万米ドルになるとの予測もされている[2]。

3.1.2. AI セーフティ評価の重要性と利活用促進への期待

このように、ヘルスケア分野においては医療機関や一般生活者向けに生成 AI プロダクト・サービスの普及が進んでいる一方で、単なる精度や利便性の観点だけではなく、「安全に活用できるかどうか」を見極める視点が不可欠である。すなわち、生成 AI が提案・出力する情報が、医療機関や一般生活者にどのように影響するかを予見し、これを制御・検証可能にする評価の枠組みの整備が求められている。

こうしたニーズ・状況を踏まえ、例えば以下のようなユースケースが想定される（図 3-1）：

- 【医療機関向け】退院時サマリ作成支援：診療記録を要約し、文書の雛形を生成
- 【医療機関向け】患者説明支援：疾患や治療方針の要点を自然言語で整理・提示
- 【一般生活者向け】健康情報の提供：PHR を踏まえた食事や運動に関する情報を提示

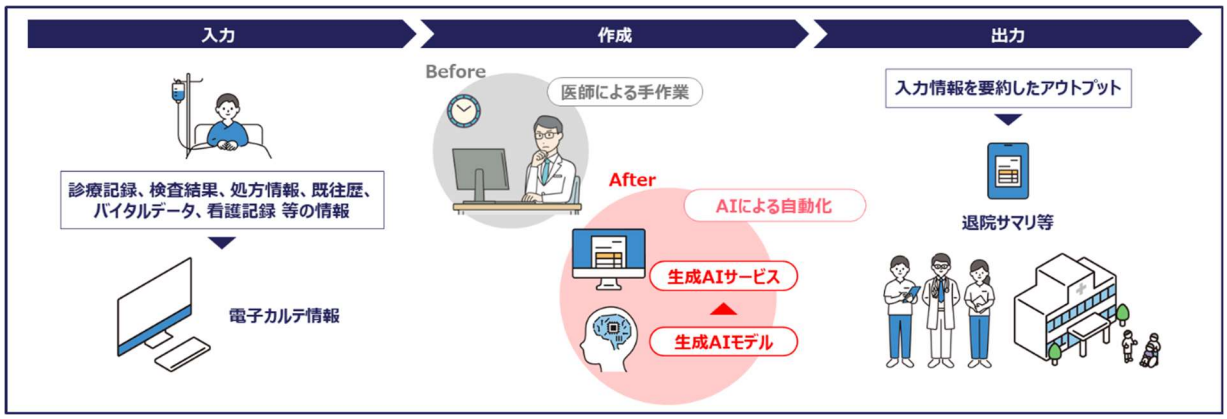

図 3-1 医療現場における生成 AI 活用の例（退院時サマリ作成等）

しかしながら、ヘルスケア分野は個々人の生命や身体に直接影響が及ぶおそれがある領域であるため、例えば生成 AI の出力結果のハルシネーションや、取り扱う情報の管理に関する懸念も根

[1] 医師の働き方改革特設サイト
https://iryou-ishi-hatarakikata.mhlw.go.jp/about/

[2] Fortune Business Insights
https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-in-healthcare-market-100534

強い。こうした懸念に対して「信頼に足る」と評価されるかどうかが生成 AI プロダクト・サービスの活用・導入可否を判断する際の重要な分岐点となっている。こうした流れを受けて、一般社団法人日本デジタルヘルス・アライアンス（JaDHA）では、『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）を策定し、ヘルスケア事業者が生成 AI を用いたプロダクト・サービスを提供する際に、利用者に不当な不利益を供することとならないよう、当該サービスを提供しようとする事業者がセルフチェックできる目安となるチェックポイントを提示している。本業界ガイドラインによって、生成 AI 活用時にどのような観点を確認しながらプロダクト・サービスを提供するべきかのメルクマールは一定程度可視化された。一方で、そうした観点がどの程度遵守できていれば「信頼に足る」と評価されるかの評価項目や評価方法が定まっておらず、このような確立した評価がない状態での導入・活用には依然として一定の懸念が示されている。これが、ヘルスケア分野における生成 AI の社会実装に向けての課題の一つとなっている。

そこで、想定されるユースケースごとに評価手法と検証事例が体系化・可視化され、プロダクト・サービスの設計段階からのセーフティの評価項目を反映することにより、生成 AI を安心・安全に利活用できる環境の整備が期待されているところである。

本 SWG では、こうした状況を踏まえ、ヘルスケア分野における生成 AI の社会実装に向けた柔軟な評価観点の整備を進めていく。（図 3-2）

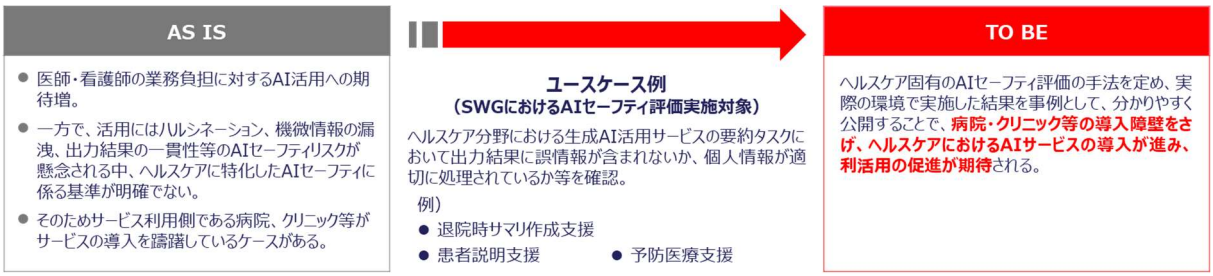

図 3-2 AS IS／ユースケース例／TO BE の対応表

3.1.3. 想定されるリスク例

ヘルスケア分野において想定されるリスクと評価観点の例は以下のとおりである。今後ユースケースに基づいて SWG にて議論・検討を進めることで、整理・精緻化を目指す予定である：

- プライバシー漏洩：個人情報を含むデータのマスキング処理が不完全なまま入出力される。
- ハルシネーション：事実誤認を含む文章の生成により、診療ミスや誤情報提供につながる。
- 出力の一貫性欠如：同一条件でも出力がばらつく場合、記録としての信頼性が損なわれる。
- バイアス混入：年齢・性別・疾患タイプなどに関する学習データの偏りにより不適切な出力がなされる。
- 責任所在の不明確さ：出力の結果に対する責任の帰属が不明瞭になる。
- 説明可能性の欠如：出力の根拠が不明瞭なことにより、医療現場での理解が困難になる。

上記の想定リスクを踏まえ、AI セーフティ評価の観点としては以下のような例が考えられる。

- 個人情報が適切に処理されているか
- ハルシネーションを低減する工夫がなされているか、重要な項目の抜け落ちがないか
- 出力結果に一貫性があるか
- 生成 AI サービスを導入する医療機関等のネットワーク環境がセキュアか

各ユースケースに、必要となる評価の観点を取り込み、定量的・定性的な評価モデルを整備するとともに、医療従事者へのユーザビリティに配慮することが、実装を進める上で不可欠である。

3.1.4. AI セーフティ評価の方向性

本 SWG では、既存の AISI の『AI セーフティに関する評価観点ガイド』や、JaDHA の『ヘルスケア事業者のための生成 AI 活用ガイド』を踏まえ、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。

また、本取組みの対象とする生成 AI による出力形式としては、テキスト、画像、音声やマルチモーダルなどが考えられるが、まずはヘルスケア領域で最も広く活用されていると考えられるテキスト生成 AI を対象とする。また、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（いわゆる Non-SaMD：Non-Software as a Medical Device）を想定した AI セーフティの評価に取り組むこととする。なお、本取組みの対象については、今後の生成 AI の発展や業界におけるサービスの進展等を踏まえて随時アップデートも想定しているところである。

評価項目の例としては以下が挙げられるが、これらについても、今後、具体的なユースケースに基づいて SWG にて議論・検討を進めることで整理・精緻化を目指す予定である：

- 出力の再現性：同一入力に対して、一定の出力一貫性が保たれているかを確認する。
- ハルシネーションの検出：虚偽情報の自動生成を未然に防ぐ仕組みが組み込まれているか、実装の有無と精度を検証する。
- プライバシー保護：特に生成 AI の入出力において、個人情報のマスキング処理が適切に実装されているかを確認する。
- 出力品質の指標化：文体、読みやすさ、誤記・要約漏れの有無など、実際に医療現場での利用に足る品質水準を定量・定性的に評価する。

これらの評価項目を実運用可能な形で導入するため、AI セーフティ評価に対応したチェックリスト形式の評価項目集や、入出力を同時に記録・比較する評価ツールの開発を進める。加えて、複数の医療機関や AI 開発・提供事業者との連携のもとで、上記評価項目を適用し、実データを用いたフィージビリティ検証に取り組む。評価結果を踏まえ、今後の評価フレームの汎用性と現場適合性の向上に資することが期待される。

3.1.5. ロードマップ

ヘルスケア SWG におけるロードマップは以下の通りである。ただし、ロードマップは検討状況や技術確認などの外部環境も踏まえながら適宜見直しを図っていく予定である。なお、ヘルスケア分野におけるユースケースは toB（例：医療機関向け）および toC（一般生活者向け）の双方が想定されることから、短期的な検討の中でどのユースケースを取り扱うのかを SWG での議論を通して検討する予定である。

令和 7 年度は、Non-SaMD を対象に、代表的ユースケースを洗い出した上で、検討対象とするユースケースの選定を行う。加えて、外部有識者へのヒアリングなども実施し、生成 AI の利活用におけるリスク構造を明確化した上で、AI セーフティ評価に必要な観点や評価シナリオを設計する。さらに、個人情報のマスキング不備やハルシネーションをはじめとした想定されるリスクに着目し、それらに対応する評価観点を整理したガイドの策定や評価データセット・評価ツールの検討を進める。

中期的には、令和 7 年度に構築した評価項目やデータセット・評価ツール等を、AI 開発・提供事業者が用いて実際に有用性を検証する効果検証段階へと移行する。この段階では、例えば、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し、策定したガイドにおける評価プロセスや評価ツールのフィージビリティの検証を進める。

将来的には、整備されたガイドや評価データセット・評価ツールについて、ヘルスケア業界において活用が推進されるような広報・浸透活動を実施していく。また、ユースケースの拡大に応じて、ガイド等について継続的なアップデートを図る予定である。（図 3-3）

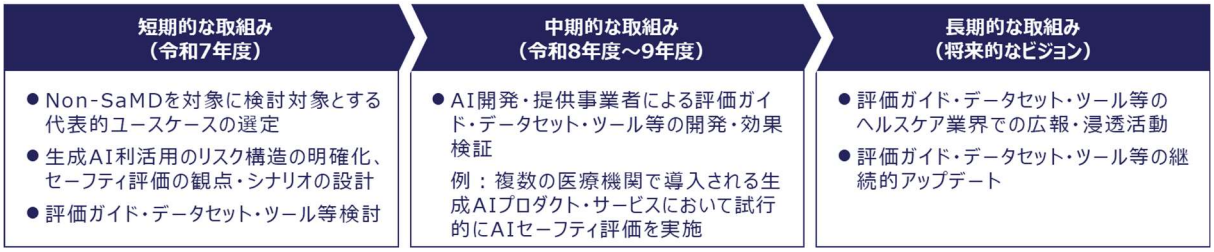

図 3-3 ヘルスケア SWG のロードマップ

3.2. ロボティクス SWG

人と空間を共有しながらサービスを提供するサービスロボット、また工場等で用いられる協働ロボットなどロボティクス分野では、AI の搭載によりその振る舞いがより複雑かつ柔軟になりつつある。対話型ロボットや自律移動型ロボットの普及に伴い、認知・判断・行動の各段階における AI の関与が拡大する一方で、その判断の妥当性や利用者への影響に対する新たな安全性の懸念が浮上している。従来の、主に物理的な危害に対する機能安全に加え、予防的なリスク低減や対人インタラクションにおける信頼形成や文脈理解、人の支援といった観点が不可欠となる中で、ロボティクスにおける従来の機能安全を含んだ新たな安全性評価、すなわち AI セーフティ評価の設計と検証が喫緊の課題である。

本節では、ロボティクスにおける AI 利活用の現状を概観しつつ、具体的なユースケースを想定した AI セーフティ評価に係る検討課題や取組み指針を取りまとめる。

3.2.1. AI 利活用の現状

近年、ロボティクス分野においても、AI の応用が急速に拡大している。特に、移動・対話・作業の 3 軸にまたがる高度な判断・制御を求められるサービスロボットや協働ロボットにおいては、センサ・アクチュエータの物理統合を越えて、「認知→判断→行動」の一連の知的処理を担う AI の役割が不可欠となっている。例えば、飲食店や空港で稼働する移動型ロボットにおいては、障害物回避や経路最適化といった従来の制御課題に加えて、利用者との自然なインタラクション、適切な指示表現や判断タイミングの調整など、文脈に応じた行動の適応性が求められる。一方、産業用ロボットでは従来ロボットの利用が難しかった食品や医薬品等の産業分野、物流搬送等の用途、中小企業等組織での利用など、いわゆるロングテール領域での活用が増えつつあり、そうした場面においてはこれまでより柔軟かつ容易に利用できるよう AI により汎用性を増した協働ロボットが求められるようになっている。

このようなニーズの高まりを背景に、ロボティクス AI 市場は拡大基調にあり、今後年間 26.8% 以上で成長し、2031 年には 941 億ドル（約 13.6 兆円）の市場規模になるとの予測もある[3]。とりわけ、生活支援ロボットやコミュニケーションロボットなど、対人インターフェースを持つ領域においては、生成 AI や視覚言語モデル（VLM）などのマルチモーダル生成 AI の導入によって、人間の五感に基づいてロボットが判断を代替する等、これまでにない利用者体験の実現が模索されている。また、協働ロボットにおいては、人間とロボットの間での協働だけではなく、ロボットがロボットを制御し協働する可能性も考えられる。近年では、「Embodied AI」と呼ばれる、人間との対話と物理環境との相互作用を統合したロボット知能の開発も進展しており、これにより AI による状況理解と行動選択がより高度に融合されつつある。（表 3-1）

[3] Statista
https://www.statista.com/outlook/tmo/artificial-intelligence/ai-robotics/worldwide

表 3-1 ロボティクス領域における AI のユースケースマップ（対話型／移動型／協働型）

	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

3.2.2. AI セーフティ評価の重要性と利活用促進への期待

ロボティクス分野における AI 利活用では、「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、AI セーフティの検討が必要となる。特に、サービスロボットや協働ロボットに代表される人と接するロボットにおいては、従来のリスク対応を担う機能安全（物理的な危害：衝突の回避や障害物検知など）に加え、対人インタラクションの質や文脈に沿った判断を評価することが必要になる。これを踏まえて、ロボティクス分野においては、AI とロボットの複合的な振る舞いが生み出す便益と、そこに内在するリスクとのバランスをいかに捉え、社会実装に適した「安全性評価」の枠組みを構築するかが重要である。

現状「AS IS」としては、例えば以下のような課題が取り沙汰されている：

- 対話型ロボット：過剰に命令的な応答や、特定属性に対する不適切な発話が発生しうる。
- 移動型ロボット：経路誤認による誤誘導や誤進入、周囲環境認識の失敗による衝突リスクが残存している。
- 協働型ロボット：制御不能による誤動作での作業ミス、重大事故の発生リスクがある。

こうした課題に対し、実際の現場シナリオに即したユースケースを特定し、安全性の観点から必要な評価項目の明確化を進める必要がある。その際、公道や公共施設等の不特定多数の利用を想定するのか、または施設内での特定の利用を想定するのか等の考慮も必要になる。

例えば、ユースケースとしては以下が挙げられる：

- 高齢者や子どもとの対話や行動における制御支援
- 横断歩道等の高リスク環境の認識と自律経路選択を行う移動型ロボット
- 飲食店内での複数ロボットや予防的な人回避による協調行動（順序・回避）

将来像「TO BE」としては、ユースケースにおける評価指針や事例集が整備・公開されることで、AI 開発・提供事業者による AI 搭載ロボットの社会実装判断が容易になり、設計・開発・運用段階における AI セーフティへの配慮の定着、ひいては、利用者の導入障壁を下げ、利活用の促進が期待される。（図 3-4）

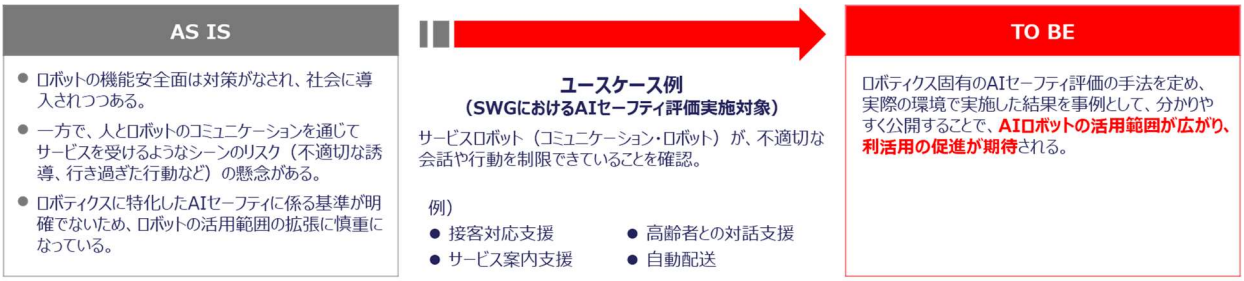

図 3-4 AS IS／ユースケース例／TO BE の対応表

3.2.3. 想定されるリスク例

ロボティクス分野の AI セーフティ評価にあたっては、物理的リスクの低減だけでなく、人間との文脈的インタラクションの中で発生しうる心理的リスクや社会的リスク等の抽出と制御が不可欠である。（表 3-2）

表 3-2 人間とのインタラクションにおける AI セーフティのリスクの例

分類	リスク	主な内容	具体例
物理的リスク	衝突・接触	ロボットの腕や車輪、体が人間に衝突・接触	産業用ロボットアームの誤動作、移動ロボットの経路ミスによる衝突
	挟み込み・切創	可動部に指や手が巻き込まれる	協働ロボットの制御不能による接近時
	感電・火災	バッテリー漏電・発火、導通異常	サービスロボットのAI学習による高負荷時の発熱
心理的リスク	恐怖・威圧	不適切な挙動	大きな音・急な動作・威圧的・意味不明な発話
	誤認・過信	ロボットの作業ミスによるストレス・損失	誤った判断・作業方法による業務への影響
社会的リスク	プライバシー侵害	会話・画像・音声などの収集・解析	会話内容の無断での保存、位置情報の追跡
	偏見・差別の助長	学習データに基づく生成AIの差別的応答	性別・人種・年齢に関する不適切発言の生成
	アクセシビリティ・公平性の欠如	利用者の属性に適さない対応	利用者が理解できない難しい説明
	信頼性低下・誤報拡散	ロボットが誤った情報を生成・提供	誤案内、フェイクニュース生成・拡散

以下は、現時点で特に重要と想定しているリスクである：

- 誤動作・誤誘導：狭小空間や段差環境におけるナビゲーション失敗や、誤認識に基づく不適切な誘導行動
- 感情的・心理的影響：利用者に対する威圧的、または意味不明な発話による不安感や不信感の醸成
- アクセシビリティと公平性の欠如：認知能力や身体的特性の異なる利用者への対応不足による、不平等なサービス提供

これらのリスクに対して、現在、AI の利用が適切ではない機能やシステム、安全開発評価プロセスの明確化や特に生成 AI やマルチモーダル AI を統合したインタラクションにおける評価枠組み（行動の文脈適合性や発話の危険度の検出指標の体系化等）が検討課題となっている。本 SWG では、人への危害の可能性などの高いリスクに対処するための、従来型の機能安全への AI の利用に関する議論や取組み（ISO/IEC における国際標準化検討）との整合性を担保しつつ、それらと一線を画した、AI を利用することにより新たに発生する可能性のあるリスクを探索、抽出し、社会への実装と受容の進展に応じた許容水準を探るとともに、これらのリスクを制御する手法を明示する新たな評価軸の確立を目指している。例えば、ロボットが利用者の意図に反した行動を抑止する能力や、特定の属性に依存しない均質な対応をどの程度実現できているかといった指標が、今後の評価の枠組みにおいて重要性を増すと考えられる。

3.2.4. AI セーフティ評価の方向性

本 SWG では、「AI ロボットがもたらす便益に対し、人は AI を利用することで新たに発生しうる、どのようなリスクを、どの程度まで受容できるか」という観点を中心に、模擬環境下での評価実証（川崎重工業株式会社が運営するソーシャルイノベーション共創拠点 KAWARUBA[4]等）および模擬シナリオを活用したシミュレーションの双方を組み合わせた評価設計を試みる。例えば、生成 AI を活用したロボットが不適切な発話を行わないようにする制御手法、特定の利用者層に対する配慮動作（例：高齢者への速度制限、視線制御）など、行動設計に直結する安全性指標の可視化等を想定している。

評価対象となる技術要素は以下の 3 カテゴリに大別される：

- センシングと認知系
  例：ビジョンセンサによる人物認識の正確性、音声認識の雑音耐性、マルチモーダル入力の統合整合性
- 移動・動作制御系
  例：移動アルゴリズムの経路安定性、障害物回避精度、環境変化（人の飛び出し等）への即応性、意図推定に基づく動作予測の確度
- インターフェースと説明性
  例：遠隔オペレーターがロボットの意図を正確に把握できる操作 UI、説明可能な判断理由の提示、利用者による意思決定支援機能の整備

これらの技術要素を基に、「ロボットの種類（対話型、移動型、協働型）×AI の機能（認知、判断、行動、人操作支援）」のマトリクスでリスクを分類し、開発、評価、運用の各プロセスにおける AI の利用、複数ロボットやインフラを含む階層システム観点などのいくつかの切り口をもとに、それぞれに適した AI セーフティ基準を策定することにより、今後の拡張的応用にも耐えうる評価の枠組みの基盤を形成する。

また、評価方法論としては、定量評価と定性評価を統合的に設計する必要があり、その相互補完性の確保が焦点になると考えられる。例えば定量評価は、不適切発言の頻度や応答時間のばらつきなど利用状況の指標化により把握される。一方、定性評価は、発話内容の印象や対話の受容性といった利用者の主観や感性に関わる要素を観察・分析する主観評価などにより把握される。これらを評価フレーム上で接続し、総合的な安心感等の指標へと昇華させることが今後の設計目標となる。

加えて、模擬環境における評価実証（例： KAWARUBA を活用した対話型ロボット実験）を通じて、出力制御や行動設計の修正ポイントを抽出する。これらの知見は、今後のロボット行動モデルの最適化や、出力制御の実装基準の策定にフィードバックされ、AI セーフティのガードレールの設定やフォールバック設計等のリスクの軽減に寄与することが期待される。（図 3-5）

本 SWG の AI セーフティ評価の取組みは、ロボットの構造やタスク特性ごとに応じた安全性評

[4] CO-CREAION PARK KAWARUBA
https://kawaruba.com/

価の体系化を進める上での出発点となるものであり、許容されうるリスクの範囲を明確にしながら、社会受容性を高める評価モデルの構築を目指す。

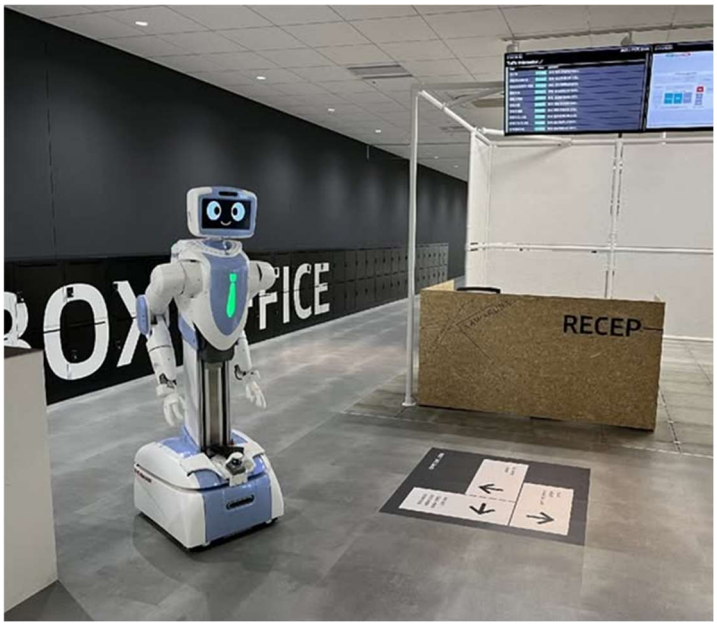

図 3-5 KAWARUBA における実証環境の概要と利用シーン例

出所）川崎重工業株式会社

3.2.5. ロードマップ

本 SWG では、まず対話型・移動型・協働型といったロボット類型ごとに、インタラクションを起点とした AI セーフティのリスクの抽出を行い、多層的な評価シナリオの設計を進める。加えて、模擬環境を活用したリスクへの対応策の検討や、初期的な試行評価を行う。

中期的には、実環境での導入効果を検証しながら、リスクに関する指標の体系化や、受容性・信頼性といった要素を含む統合的分析手法の確立・高度化に取り組む。

将来的には、社会実装の拡大に向けて、評価モデルや設計指針の体系化と共有を進め、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図る。また、運用段階での継続的な評価と改善を可能とする仕組みを整備し、実環境での信頼性を持続的に確保できる体制の構築を目指す。（図 3-6）

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）
● 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	● 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	● 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

図 3-6 ロボティクス SWG のロードマップ

3.3. データ品質 SWG

AI システムの信頼性を確保する上で、学習・評価・活用といった各フェーズで使用されるデータの品質は極めて重要な基盤である。AI は「Garbage In, Garbage Out」の構造を持ち、特に、非定型な出力を持つ生成 AI では、入力データの欠陥が予期しない出力につながりうることから、データ品質の可視化と管理手法の確立と実践が課題となっている。一方で、データ品質にかかわる国際標準は多様で、すべてを考慮した実装は難しいという指摘がある。また、それらの管理対象は基幹系システムや構造化データが中心であり、テキストや画像など非構造化データ、動的に参照されるナレッジベース、第三者提供のデータなど、管理対象となるデータの範囲と複雑性が飛躍的に拡大する AI 時代のデータ品質をカバーしきれていない課題も浮上している。

本節では、AI セーフティの実現に向けて不可欠なデータ品質マネジメントについて、国内外の標準化動向を踏まえながら、実務への展開を意識した支援のあり方に焦点を当てる。具体的には、評価観点の整理、チェックリストやスコアリングツールの整備、現場での適用を通じたフィードバックの収集と改善といった取組みを中心に、その方向性について述べる。

3.3.1. 現状と課題

AI は本質的にデータ・ドリブンな技術であり、その信頼性は学習・評価・活用に用いるデータの品質が直結する、これは AI セーフティの前提条件であり、『AI セーフティに関する評価観点ガイド』においても、10 項目の評価観点の一つとしてデータ品質が取り上げられている。

従来の AI では学習・評価データの品質管理が中心であり、AI 開発・提供事業者には関心のある領域であった。生成 AI の普及に伴い、AI の開発・提供から利用へと広がり、さらに RAG や AI エージェントといったアーキテクチャの進展により、組織内外のデータの動的な参照が一般化しつつある。このような利用形態は、AI が学習した知識を補完し、ハルシネーションを防ぐ仕組みとして期待される一方で、参照するデータの正確性や最新性が欠けていれば、かえって誤った出力を助長しかねない。したがって、学習データと同様に、参照するデータの品質とその来歴の管理も重要となっている。

データ品質については複数の国際標準が提唱されている。代表的な標準としては、ISO 8000（データ品質管理全般）、ISO/IEC 25012（ソフトウェアにおけるデータ品質特性）、ISO/IEC 5259 シリーズ（AI システムにおけるデータ品質）があり、理論的な枠組みとしての整備は進んでいる。日本国内では AISI が 2025 年 3 月に実務の点で整理した『データ品質マネジメントガイドブック Ver.1.0』[6]を公表し、国際標準と実務の間をつなぐ取組みも始まっている。

一方で、以下のような技術的・構造的課題が未解決である：

- 標準と実務のギャップ
  現在、データ品質に関する標準類は複数存在するが、いずれも定義が中心であり、実務における具体的な評価方法や導入手順までをカバーしているわけではない。例えば、データの正確性、偏りといった概念を、評価可能な指標に落とし込む手法や評価のタイミングが確立されていない。結果として、「どの標準を参照すればよいか分からない」「自組織に合

った適用方法が見えない」といった実装上の障壁が多くの現場で指摘されている。また、標準が多い中で、過度なコストや過度なガバナンスを避けたいという現場の声も多く、必要十分でリーズナブルなデータ品質管理が求められている。

- AI 時代のデータ品質マネジメント

  標準策定には時間がかかるため、生成 AI や RAG のような新たなデータ利用形態には十分に対応できないことがある。例えば、「テキストや画像などの非構造化データの品質評価において、従来の統計手法では限界がある」「RAG で参照する外部データの品質評価ができない」「機械にとっての可読性と人間にとっての可読性の実践的なあり方が不明」等の課題がある。このように、扱うデータの拡大に伴う品質評価が属人的判断に頼らざるを得ない状況が続いている。

- データ品質評価への関心

  データ品質の重要性自体が経営層や組織内で十分に認識されておらず、取組みの優先度が上がらないという課題も存在する。これは、データという目に見えない対象を扱う性質に加え、データそのものは最終的なアウトカムに対して中間的な位置づけにあるため、AI モデルや分析結果といった成果物の品質に注目が集まりやすいことが要因である。しかしながら、信頼できる AI を実現する上で、データこそが安全を支える基盤であるという認識を高めることが不可欠である。

3.3.2. データ品質の確保に向けた取組み方針

データ品質管理は、『データ品質マネジメントガイドブック Ver.1.0』で説明されたとおり、データのライフサイクル全体にわたる 8 つの工程（計画・収集・準備・加工・AI 処理・評価・提供・廃棄）を対象に構成し、各工程における品質評価ポイントの体系化により実現する（図 3-7）。つまり、いつ、だれが、何を、どう評価すればよいかを整理し、最終的には第三者へのデータ流通も含めて信頼のチェーンを築くことである。さらに、ライフサイクル全体にかかるデータマネジメントおよびデータガバナンスによってデータ品質を確保するアプローチをとる。

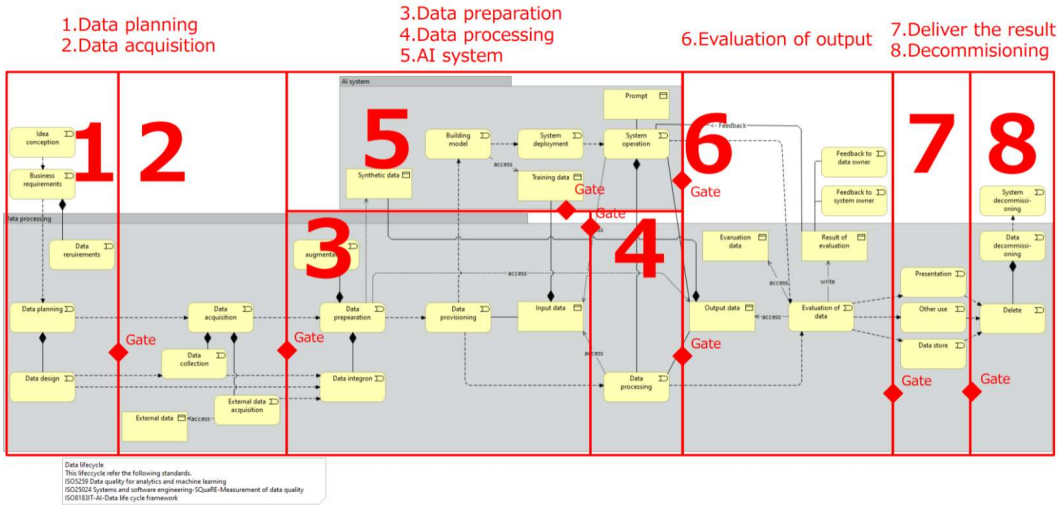

図 3-7 AI システムも含めたデータライフサイクル

本 SWG では実務志向で、データ品質管理の実装障壁を下げ、組織での実践を促進することを目的として、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の整備に取り組む。また、基幹系システムにおける定型データ品質、機械学習における教師データ品質、生成 AI 時代の参照データ品質といった異なる文脈にあるデータ品質を統合的に捉えたマネジメントのフレームワークを提示し、これを実務へ橋渡ししていくことを目指す。

その実現に向けては、データ品質管理に精通した AI 開発・提供事業者や研究者など、様々な知見を集めるとともに、現場における適用検証を通じて実証的に活動する。

図 3-8 データ品質 SWG の取組み方針

主な取組みは以下のとおりである：

- ガイドの整備

  既存のガイドラインや標準類を参照・補完しつつ、工程別・データ形式別に具体的な評価観点と実践フレームワークを整理したガイドブックをアップデートする。また、入門版や、実装イメージを喚起するユースケースなどもあわせて整備する。

- 評価ツールの提供

  AI 開発・提供事業者がデータ品質マネジメントの第一歩として活用できる評価ツールを提供し、実践障壁の低減を図る。以下は検討中のツール実装例である。

  - 評価対象データ：学習データ、セーフティ評価データ、参照データなど
  - 定量評価：データの統計分布、欠損率、偏りの指標化など
  - 定性評価：メタデータの有無、品質マネジメント実施状況など
  - ツール実装：チェックシート、簡易診断ツール、自動評価スクリプト（Python アプリ等の軽量な実装を想定）。特別な計算環境や高度な技術スキルを不要とする
  - 出力フォーマット：レーダーチャートやスコア形式など、直感的に可視化・理解可能な形式で提供し、実運用者が活用しやすいよう設計する。

  将来的には高度な評価処理、自動化、継続性のあり方も検討する。

- 適用検証の実施とフィードバックの反映

分野別 SWG 等と連携し、実際の業務シナリオに即したデータ品質評価のトライアルを実施する。評価結果および現場からのフィードバックを踏まえ、ガイドやツールの改善を行う。また、適用事例や導入時の課題を体系的に収集・分析することで、他分野・他組織への展開を支援し、標準と現場とのギャップを段階的に埋めていく。

これらの取組みを通じて、多様なデータ形式や活用シーンに対応可能な柔軟な品質評価フレームワークを確立し、データ品質の観点で AI セーフティの社会実装を支える基盤の構築を進める。

まずは評価観点や出力形式についてニーズと実現可能性に応じた優先順位を定め、ガイドの継続的な改善を図る。評価ツールについては AI セーフティ評価における「データ品質」観点を具体化・補強する役割を担うものとして、分野別 SWG 等における検証シナリオにも活用していく。

3.3.3. ロードマップ

令和 7 年度は、『データ品質マネジメントガイドブック Ver.1.0』に基づく適用検証を実施し、現場のフィードバックを収集・反映し、実装の詳細化と事例を拡充する。あわせて簡易的な品質評価ツール（チェックリスト等）を試作し、分野別 SWG 等との連携の下、試行的な評価を通じて導入容易性および実効性を検証する。

中期的には、複数分野での横断的導入実証を行い、行政・産業界における活用モデルを確立するとともに、マルチモーダル等のデータ形式やマルチエージェント等の新たな利用形態にも対応するよう、ガイドや評価ツールの対象の拡充を図る。また、評価ツール機能の本格化として、分野別テンプレートや非専門家向けのユーザーインターフェース等の整備、継続的なデータ品質評価の枠組みの検討を進め、その実証的な検証に着手する。

将来的には、組織内での継続的なデータ品質評価の運用モデルを確立し、社会インフラとしての定着を目指す。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も構築する。

さらに、ISO/IEC JTC1 SC42 など国際標準化団体との連携を通じて、日本発のデータ品質評価フレームワークや指標定義を国際的な議論へ反映する。（図 3-9）

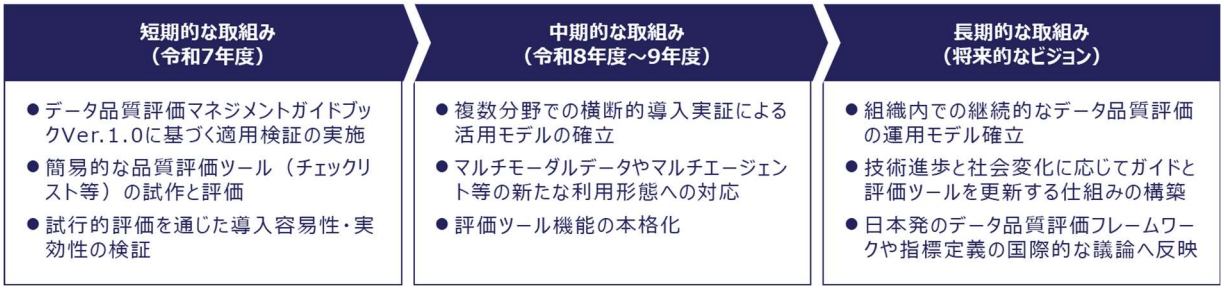

図 3-9 データ品質 SWG のロードマップ

3.4. 適合性評価 SWG

AI がサービス形態や運用体制とともに動的に変化し続ける中、従来の静的かつ一括的な適合性評価の枠組みでは対応が困難となりつつある。AI 利活用にあたり、組織がどのような取組みをしたらよいのかについて、AI マネジメントシステムの国際標準 ISO/IEC 42001 が発行された。これは、ISO9001 や ISO/IEC 27001 などと同様に組織のマネジメントシステムに関する要求事項が記載されており、組織が AI マネジメントシステムを構築する際の指針となる。また、ISO/IEC 42001 は適合性評価の基準としても第三者適合性評価制度の整備が進んでいる。

AI システムに対する適合性評価に関する国際標準としては、ハイレベルなフレームワークを規定するものではあるが、ISO/IEC 42007 の開発も始まっている。この標準は、適合性評価スキームに関する国際標準である ISO/IEC 17067 の改定と連動する形で AI システムに対する柔軟な適合性評価スキームに関するものである。

本節では、こうした背景を踏まえるとともに、①国際標準の内容を紐解き、それに基づいて組織に実装するには一工夫が必要であること、②国際標準を基礎とした適合性評価を実際に運用できるようにするためには具体化が必要であることから、フィージビリティスタディを通じて、AI における適合性評価の具体的な手法の開発、制度的・運用的システム、プロセスの構築に向け、論点を整理する。また、AI システムに対する適合性評価に関係する機関（評価側）および適合性評価を受ける AI 開発・提供事業者等の国内関係者（受審側）との連携によるパイロット評価や、将来的な実装に向けたロードマップについて述べる。

3.4.1. 現状と課題

AI 技術は日々変化しており、従来の製品・人・組織単位の静的かつ縦割りである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課題がある。このような課題意識を受け、ISO/IEC JTC1 SC42 では、AI システムに対する適合性評価のハイレベル・フレームワークを規定する ISO/IEC 42007 の国際標準化が進行中である。一方で、EU AI Act における適合性評価制度との関係など明確になっていない部分も存在する。

我が国においても、ISO/IEC 42001 や ISO/IEC 42007 を実際に活用するためには、少なくともいくつかの段階で具体化が必要となる。したがって、国内 AI 産業の活性化に資する実用的な適合性評価手法の開発について検討を進める必要がある。

ISO/IEC 42007 は、適合性評価のハイレベル枠組みを規定するものである。そのため、具体的な運用手順や評価方法の開発を考える場合には、国際標準を解釈し、実行可能なレベルにまで詳細化、具体化をしなければならず、そこに向けた検討を関係者と行う必要がある。評価方法に関しては、第一者適合性評価（自己適合）から第三者適合性評価までを含む柔軟かつ適切な評価方法と実行可能な手順となるように分析、検討を進める必要がある。

評価側の視点

- AI のような日々進歩する技術を適合性評価の対象とする場合、従来の評価方法の枠組みでは追随できない可能性が排除できない。そのため、変化する評価対象に対応した評価手

法の開発、評価粒度や妥当性基準について再検討するとともに、新たな適合性評価の手法を開発し、枠組みを再設計する必要がある。

- AI のような進化的技術に対し、従来の適合性評価制度（製品・人・組織といった縦割り）では対応が十分であるか不透明な部分がある。特にデジタル（システム・サービス）に対する適合性評価の定義や手法が発展段階であり、求められるエビデンスの（特定）提示も含む評価手法の開発が技術的に未成熟であるため、AI セーフティに対応可能なアジャイル型も考慮した柔軟なフレームの開発（例：DevOps 型ガバナンス）の必要がある。
- 上記解決に向けた評価目的に応じたエビデンス設計や論証構造型評価（Evidence-based Conformance Structure）も含む適合性評価手法の開発の必要性。

受審側の視点

- 自社の AI システムが国際標準など国際的な基準に基づき適合していることを実証するための枠組みが明確でなく、評価を受けるにあたり、その負担・コストの見通しが不透明。
- 一方的な適合性評価ではない、自己適合も含む形での現実的な適合性評価制度への期待。
  - ➢ 過度なコストをかけず、開発、提供する AI システムの適合性証明ができる柔軟な制度（自己適合も含む）が求められる。

国際的な連携に向けた視点

- 相互運用性・国際連携：グローバルでの相互承認・相互運用を見据えるには、国際標準と整合しつつ、柔軟かつ段階的な制度設計が求められる。
- 関係者との連携構造：制度設計には、評価機関、AI 開発・提供事業者、（法制化するのであれば）政策担当部局など多様なステークホルダーを巻き込む運用体制の確立が前提。

本 SWG では、これらの課題を踏まえ、実運用に耐えうる適合性評価手法の開発、AI システムのレベルに適した選択肢を整理した評価手法・手続きガイドや支援ツール設計の方針に反映する。

3.4.2. 適合性評価手法の確立への取組み方針

本 SWG では評価側・受審側組織とともに、ISO/IEC JTC1 SC42 での活動や AISI での取組みを踏まえて具体的な適合性評価手法を開発するにあたり、AISI ガイドを踏まえて開発される評価ツール等を使ったヘルスケア SWG およびロボティクス SWG の具体的な取組みなどから、AI システムに関する要求事項等を導出し、自己適合性評価を含む 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を目指す。

本 SWG では、適合性評価手法の確立とその実装や将来的な制度化への足掛かりに資するため、以下の具体的ステップに取り組む：

- ヒアリング・課題抽出：既存の制度・運用課題を洗い出し、評価側・受審側双方のニーズを整理。
- 制度設計に向けた検討：ユースケースに応じた柔軟な論証構造型適合性評価の試案を提示し、第三者評価／自己適合を含む段階的導入パターンを想定。

- ツール・体制準備：データ品質 SWG と連携し、出力整合性や安全性評価観点の統合的活用を検討。
- 国際標準との整合：ISO/IEC JTC1 SC42 等の国際的動向を踏まえ、評価ガイドの整備や国内に反映。

適合性評価の枠組みとして「論証構造型アプローチ（Argumentation-based Approach）」の導入も検討する。これは、安全性を含む特定の評価ゴールごとに、段階的かつエビデンスベースで適合性を証明していく手法である。この手法は、AI のような変化し続ける対象に対応可能である。

また、データ品質 SWG との連携により、AI の出力の整合性や、利用文脈を含めた複合的な評価方法も検討する。（図 3-10）

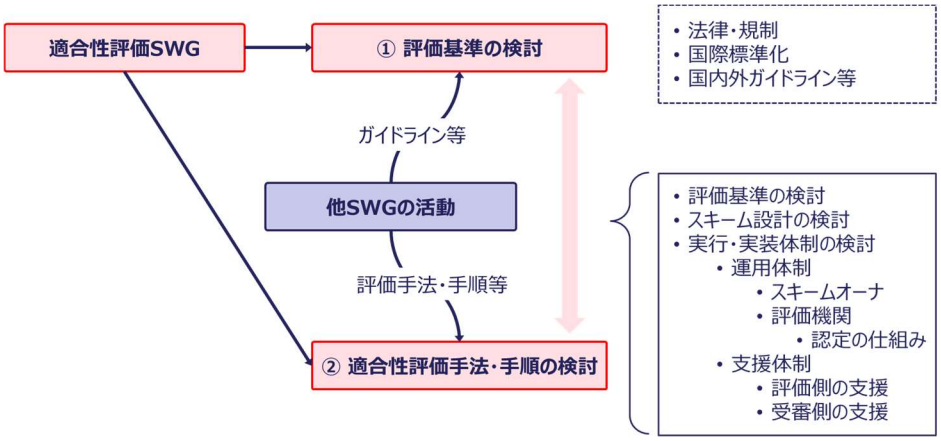

図 3-10 適合性評価手法の確立への取組み方針

3.4.3. ロードマップ

令和 7 年度は、分野別 SWG と連携し、評価ニーズや受審組織の現状に関するヒアリングを実施する。また、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の分析を進めるとともに、次年度以降を見据えた受審側の構成メンバを検討する。

中期的には、ISO/IEC 42007 を含む ISO/IEC JTC1 SC42 における国際標準化議論を踏まえて、包括的なアシュアランスおよびアカウンタビリティ確保に係る適合性評価制度の設計に向けた活動を行う。あわせて、国内における AI に関する適合性評価制度のあり方について具体化と試行にも着手し、社会実装に向けた実効性を検証する段階に移行する。

将来的には、国際的な実装モデルとの整合性（国際的な相互運用性）を踏まえたガイドライン整備を進めるとともに、国際的に相互運用可能な実運用体制（国内・海外の連携）の構築をも視野に入れた段階的な展開を図る。（図 3-11）

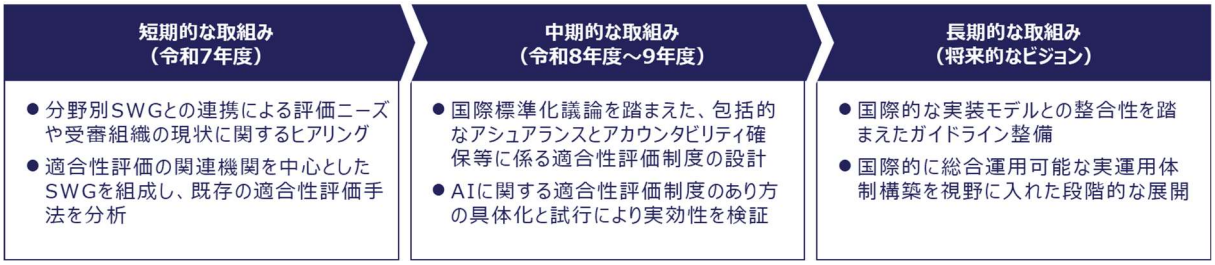

図 3-11 適合性評価 SWG のロードマップ

4. おわりに

AI 利活用は、社会課題の解決や産業の変革を支えるイノベーションの基盤として、その重要性が増している。こうした動きが加速する中、AI を安全かつ信頼して導入・利活用するためには、リスクを適切に把握し、評価し、必要な策を講じる一連のプロセスを確立することが不可欠である。

本ビジョンペーパーで示した取組みの方向性に基づいて、分野ごとのユースケースに即した実証や、分野横断的かつ国際的な整合性を意識した評価の枠組みの構築をすることにより、AI セーフティ評価の実効性が高まることを期待する。事業実証 WG は、AI の円滑な導入と利活用促進を目指して、AI セーフティ評価を推進する場を提供することにより、社会・産業の持続的成長に寄与していく所存である。

5. 参考文献

[1] 総務省／経済産業省『AI 事業者ガイドライン』（2025 年 3 月 第 1.1 版公表）
https://www.soumu.go.jp/main_sosiki/kenkyu/ai_network/02ryutsu20_04000019.html
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240419_report.html

[2] AISI『AI セーフティに関する評価観点ガイド』（2025 年 3 月 第 1.1 版公表）
https://aisi.go.jp/output/output_framework/guide_to_evaluation_perspective_on_ai_safety/

[3] AISI『AI セーフティに関するレッドチーミング手法ガイド』（2025 年 3 月 第 1.1 版公表）
https://aisi.go.jp/output/output_framework/guide_to_red_teaming_methodology_on_ai_safety/

[4] 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）
https://jadha.jp/news/news20250207.html

[5] 一般社団法人 金融データ活用推進協会『FDUA 生成 AI ガイドライン』（2024 年 10 月第 1.0 版公表）
https://www.fdua.org/news/news_20240828

[6] AISI『データ品質マネジメントガイドブック』（2025 年 3 月 Ver.1.0 公表）
https://aisi.go.jp/output/output_information/250331_2/

更新履歴

Ver	更新日付	更新内容
1.0	令和 7 年（2025 年）6 月 30 日	－
1.01	令和 7 年（2025 年）7 月 10 日	一部記載を修正