

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 8 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

目次

0. 本書の位置付け	1
1. はじめに	2
2. 事業実証 WG のビジョン	3
2.1. AI 利活用の現状	3
2.2. AI セーフティ評価の必要性	4
2.3. 政策対応と分野別の取組み	4
2.3.1. 関連する政策・規制動向	4
2.3.2. 分野ごとの取組み	5
2.4. 本事業の目的と実施体制	6
2.4.1. 目的・ゴール	6
2.4.2. 事業実証 WG の実施体制	7
2.4.3. ロードマップ	9
3. 分野別 SWG のビジョン	11
3.1. ヘルスケア SWG	11
3.1.1. AI 利活用の現状	11
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	12
3.1.3. 想定されるリスク例	13
3.1.4. AI セーフティ評価の方向性	14
3.1.5. ロードマップ	15
3.2. ロボティクス SWG	16
3.2.1. AI 利活用の現状	16
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	17
3.2.3. 想定されるリスク例	18
3.2.4. AI セーフティ評価の方向性	19
3.2.5. ロードマップ	20
3.3. 教育 SWG	22
3.3.1. AI 利活用の現状	22
3.3.2. AI セーフティ評価の重要性と利活用促進への期待	23
3.3.3. 想定されるリスク例	23
3.3.4. AI セーフティ評価の方向性	24
3.3.5. ロードマップ	24
3.4. ジオ AI SWG	25
3.4.1. AI 利活用の現状	25
3.4.2. AI セーフティ評価の重要性と利活用促進への期待	25
3.4.3. 想定されるリスク例	26
3.4.4. AI セーフティ評価の方向性	27
3.4.5. ロードマップ	28

4.	分野横断 SWG のビジョン	29
4.1.	データ品質 SWG	29
4.1.1.	現状と課題	29
4.1.2.	データ品質の確保に向けた取組み方針	31
4.1.3.	ロードマップ	33
4.2.	適合性評価 SWG	35
4.2.1.	現状と課題	35
4.2.2.	適合性評価手法の確立への取組み方針	36
4.2.3.	ロードマップ	37
5.	おわりに	39
6.	参考文献	40

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』に記載されているとおり、「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」及び「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。テキストに加え、画像、音声など様々なモダリティのデータを統合的に扱う AI モデルは、多様なアプリケーションの共通基盤として機能しつつあり、AI は既に社会・経済活動の様々な場面に取り込まれている。我が国では、少子高齢化に伴う労働力不足を背景に、その利活用は社会・産業の持続的成長に不可欠なものとなっている。

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) を設立した。AISI では、AI システムの開発者や提供者が AI セーフティ評価を実施する際に参照できる基本的な考え方を提示する各種ガイドと、それらのガイドに基づいて AI システムの評価を行うための評価ツールを順次公表している。2025 年 3 月には、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)」を設置した。

2025 年 9 月には、AI の研究開発・活用の推進に関する基本法として AI 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) が全面施行され、同年 12 月に AI 法に基づく『人工知能基本計画』及び適正性確保に関する指針が整備された。2026 年 7 月には現行の『人工知能基本計画』が閣議決定されるなど、AI 政策の基盤が段階的に整いつつある。

事業実証 WG の活動は、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』、AI に関する国際標準動向を踏まえつつ、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の整備を進めている。その活動は、制度動向とも連動しながら、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。そのため、民間事業者を中心とした多様な主体の参画のもと、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、実務的な視点からユースケースやニーズを発掘し、分野ごとの知見を踏まえた評価の枠組みを整備する。また、分野横断的な評価手法やデータ品質・適合性評価の整備により、日本発の実践的な評価枠組みの構築と、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であり、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、評価手法の確立が一層重要になっている。

本章では、こうした背景を踏まえ、「信頼できる 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、とりわけ生成 AI は、日常的な利用から業務支援まで幅広く拡大している。自然言語生成や要約、翻訳等に加え、外部情報の活用、複数のモダリティの処理、自律的なタスク遂行など、より高度な情報処理や判断支援を実現する応用も進展している。

生成 AI の重要な応用技術の一つとして、検索拡張生成（Retrieval-Augmented Generation: RAG）が挙げられる。これは、AI がリアルタイムで組織内のドキュメントやデータベース、必要に応じて外部の情報源などから関連する情報を検索し、その結果に基づいて出力を生成する仕組みであり、社内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いる。また、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、出力の生成を行うマルチモーダル生成 AI は、医用画像に基づく説明文の生成や、各種センサの情報を活用するロボットなど、実環境に即した活用が期待されている。さらに、AI エージェントの進展も著しく、予め定められた手順や業務フローに従って処理を行うものから、外部環境の変化を踏まえて自律的に選択・実行するものまで、様々な形態がある。定型的な業務プロセスの自動化に加え、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。

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付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。実際に、医療、製造、流通・接客など実世界の応用領域へと急速に広がっている。

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、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への対応も現実のものとなりつつある。これらの動きは、AI が社会・経済活動に広く組み込まれつつあ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を踏まえ、『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。イノベーションの促進と「信頼できる AI」の利活用の両立を目指し、具体的なユースケースに即した評価・実証の段階へ進みつつある。

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、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、個人情報の漏えいといった問題が現実には発生している。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 の出力には、事実に基づかない内容や根拠が不十分な内容をもっともらしく生成する「ハルシネーション」が生じる場合がある。モデルの性質や学習方法、学習データの品質などに起因するハルシネーションのリスクを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、技術的な対応をはじめとする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。

AI セーフティに関する評価観点や手法の具体化が進む一方で、AI の用途、利用環境、影響の大きさ等を踏まえて、講じた対策が十分であるか、残存するリスクが許容可能な水準にあるかをどのように評価するかについては、引き続き実務上の課題が残る。AI が社会に広く受容されるためには、可能な限りリスクの回避・低減を図った上で、残存するリスクを適切に管理することが重要であり、具体的なユースケースでの評価・検証を通じて、実践的な AI セーフティ評価の枠組みを確立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、国や地域によって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。

我が国においては、総務省及び経済産業省により『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示された。これを受け、AISI は、『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。

さらに、AI 法の全面施行、同法に基づく『人工知能基本計画』及び適正性確保に関する指針の整備に加え、国民や事業者等の能動的な取組みを促すための AISI の抜本的強化の方針が示された。事業実証 WG では、こうした制度動向を踏まえ、分野別サブワーキンググループ

(SWG)におけ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る。なお、ヘルスケアやロボティクスなど、既に法律や業界ガイドラインが存在する分野では、それらとの整合性・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 によりリスクに応じた義務の層別化（例：高リスク AI に対するリスク管理、技術文書、適合性評価等の義務）を制度化する一方、米国は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）等を通じて、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を進めている。

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。

こうした国内外の制度・標準化動向を踏まえ、多様な産業や企業に実装可能で、かつ国際標準との相互運用性を確保し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。

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みも進展しており、『AI 事業者ガイドライン』が掲げる「事業者の自主的な取組の支援」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）が生成 AI を活用したサービス・プロダクトを提供する事業者向けの自主ガイドライン^[4]を策定するなど、実務的な手引きとして提示している。また、メンタルヘルスケアのようなセンシティブな領域においても、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ている^[5]。

ロボティクス分野でも、日本ロボット工業会（JARA）が公表した『ロボット産業ビジョン 2050』^[6]において、AI の判断に基づく行動による事故の際の責任の所在や遠隔操作による悪用などが課題として示されているほか、AI ロボットの安全性評価に関する民間での検討も進められている。ロボティクスを含む先端技術の健全な普及には、ルールメイキングとガバナンスの観点を含めた長期ビジョンの策定が重要であるとされている。

このほか、金融分野では、生成 AI の適正利用に向けたガイドラインの整備^[7]や内部ガバナンス体制の強化、社内ルールの整備が進められている。教育、製造、自治体等においても、各現場の特性に応じたガイドラインの策定や評価観点といった取組みが見られる。このように、業種・業態を問わず、各分野の特性に応じた AI セーフティの確保の取組みが広がりつつある。

これらの民間主導の取組みは、AISI が示す AI セーフティ評価の共通的な観点を、分野ごとの特性や利用文脈に応じて具体化するものと位置付けられる。事業実証 WG では、既存の民間ガイドライン等の知見も踏まえつつ、分野別 SWG での実証と検証を通じて、各分野における実装可能な評価手法の具体化・高度化を進めている。

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、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、「信頼できる AI」を実現するための AI システムに求められる要件を明確にし、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、及び 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、「信頼できる AI」の利活用と実装を加速させる枠組みである。（図 2-1）

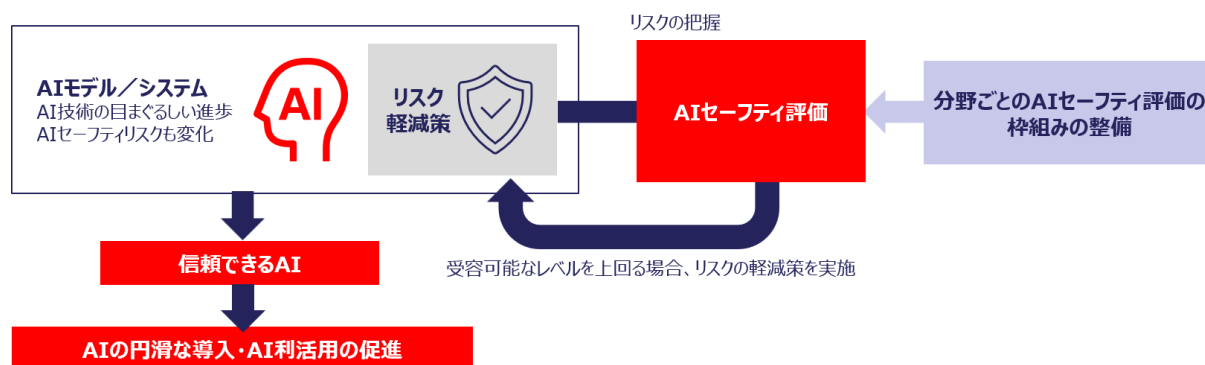


図 2-1 事業実証 WG の活動の全体像

2.4.2. 事業実証 WG の実施体制

事業実証 WG では、「信頼できる 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、各分野向けの評価観点ガイド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 では、国際的にも関心が高く、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、ロボティクス、教育、ジオ AI を対象に、AI セーフティ評価に関する各分野の特性や SWG ごとの進捗に応じて評価観点・評価手法の検討を進め、事業実証へとつなげていく。

分野横断 SWG では、分野別・具体的ユースケースごとの検討を支える観点から、データ品質や適合性評価など分野共通の評価手法を検討し、分野別 SWG の取組みと相互補完的に機能する AI セーフティ評価の枠組みを整備する。

（1）分野別 SWG

■ 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、多様なサービスが生まれ、各々の領域で課題解決に貢献する事例が登場するなど 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

考え方を踏まえた 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。

■ 教育 SWG：

教育分野は、授業設計や教材作成、校務・教務の効率化、個別最適な学習への支援、特別支援、多文化・多言語支援のように多様な児童生徒への対応など、様々な場面で AI の利活用が期待されており、教育分野に特化した AI サービスの開発・実装や教育課題の解決に向けた実証が進みつつある。一方で、年齢や発達の段階に応じた配慮、学習内容・学習機会への影響、個人情報や学習履歴等のデータの取扱いなど、教育現場に固有のリスクも想定される。そのため、教育分野のユースケースを踏まえた AI セーフティ評価項目や方法を明確化し、児童生徒や学生等の学習者、教職員向けの生成 AI サービスを安心・安全に利活用するための評価の観点を整理することが期待されている。

■ ジオ AI SWG：

ジオ AI 分野では、地理空間情報（Geospatial Information）の整備・更新に AI を活用する取り組みや、地理空間情報を AI による分析・推論・意思決定支援に活用する取り組みが進んでいる。一方で、AI を活用して生成・更新される地理空間情報の品質・信頼性や、空間関係の適切な解釈、プライバシー・セキュリティ、不適切な推論・利用等のリスクも想定される。そのため、ジオ AI のユースケースを踏まえ、必要に応じてデータ品質 SWG 等の分野横断 SWG と連携しながら、地理空間情報等を用いた AI 利活用における AI セーフティ評価の観点を整理することが期待されている。

(2) 分野横断 SWG

■ 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、その成果を継続的に蓄積・高度化していく。また、今後にも必要に応じて 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

価値ニーズが抽出され、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な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 2-2)

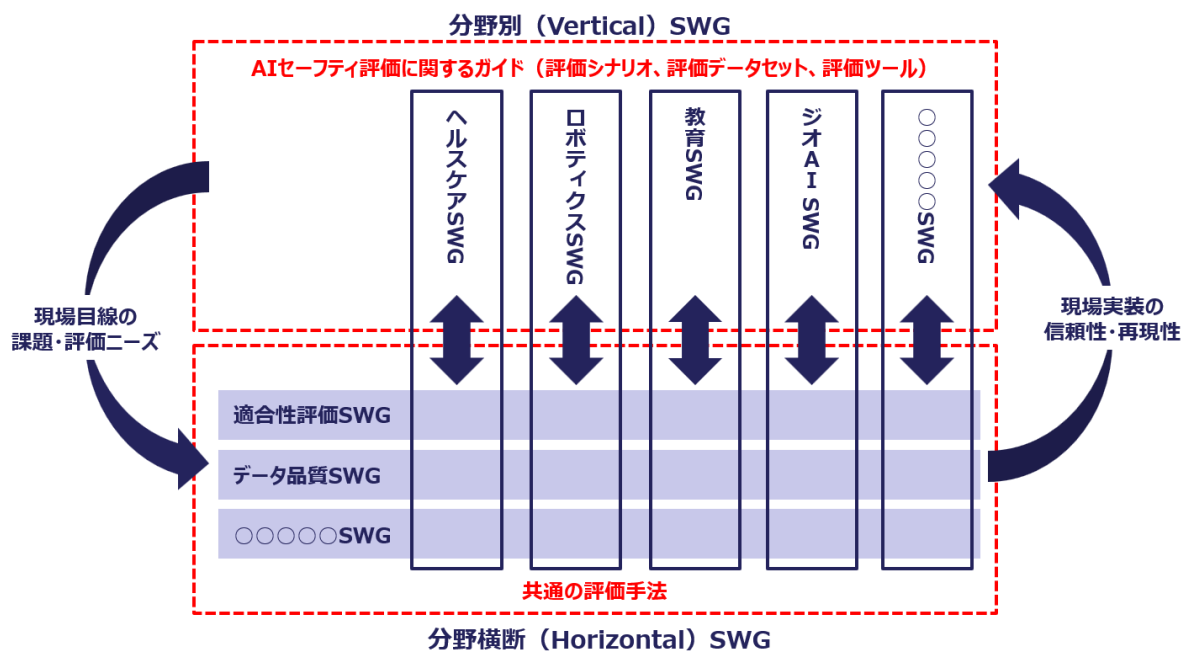


図 2-2 分野別 (Vertical) /分野横断 (Horizontal) の俯瞰構造図

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り組みを進めていく。

2025 年度は、各 SWG の活動を通じて、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、AI セーフティ評価を実務に落とし込むための評価の枠組みを示した評価ガイド (指針) を策定した。また、分野横断的に共通するリスク要素も抽出し、実装現場における評価の適用に向けた評価環境の形成に向けた初期的な議論を行った。

2026 年度以降、中期的には、評価ガイドを基盤として、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、AI エージェントシステム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 2-3）

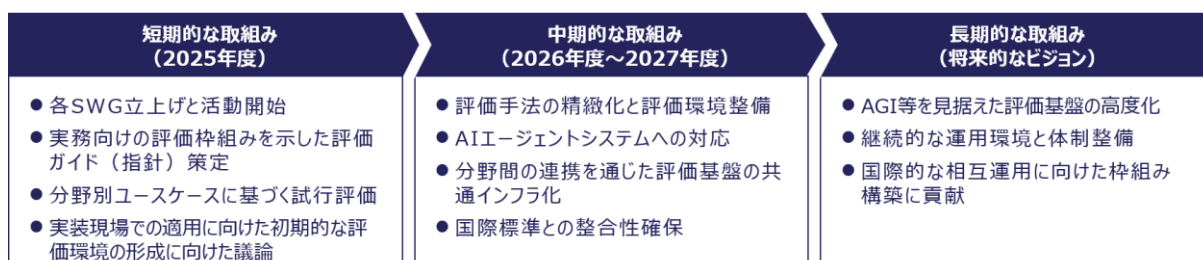


図 2-3 事業実証 WG のロードマップ

3. 分野別 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、分野別 SWG 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十分に確立されていない。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・教育・ジオ AI といった各領域の進展を踏まえ、各分野のユースケースに即した評価の観点や方法を具体化し、AI セーフティ評価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 に示したとおり、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 月から開始された医師に対する時間外・休日労働の上限規制の適用を背景として、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により多くの時間を充てることを可能にするものとして、大きな注目を集めている。

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 を用いて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。

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

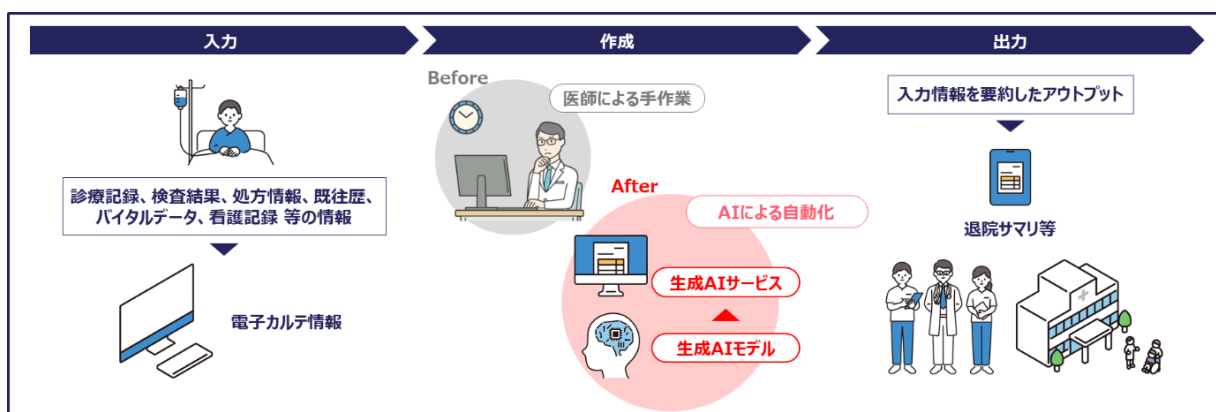


図 3-1 医療現場における生成 AI 活用の例（退院時サマリ作成等）

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強い。こうした懸念に対して「信頼できる AI」と評価されるかどうかが生生成 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』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生じさせることがないように、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「信頼できる AI」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）

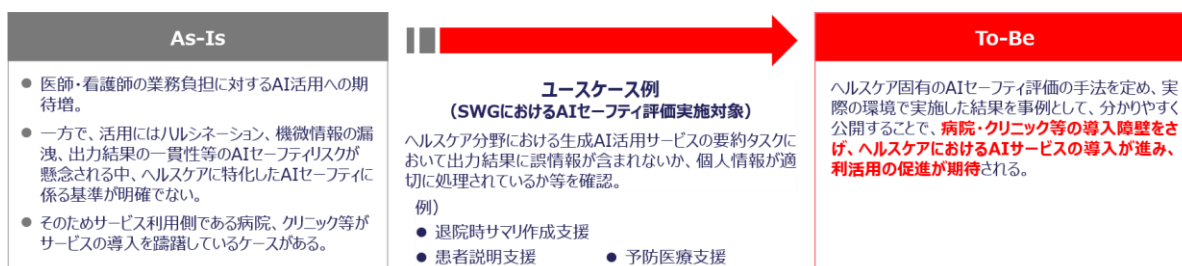


図 3-2 As-Is / ユースケース例 / To-Be の対応表

3.1.3. 想定されるリスク例

ヘルスケア分野における生成 AI 活用では、AI セーフティの観点から、特に慎重な検討を要する以下のようなリスクが存在する。

- 患者の生命・身体・精神への直接的影響

ヘルスケア領域では、医療機関や生活者に対して健康に関する情報を提供する場面が想定されるため、不正確な出力や不適切な応答がユーザーに対して直接影響を与える可能性がある。

- 取り扱うデータの機微性

ヘルスケア領域では個人の症状に関する情報やバイタルデータから普段の生活に関する情報まで、健康・疾患に関する幅広いデータを扱うことが想定される。特に、遺伝情報や既往歴・健康診断結果に関する情報など、個人情報の中でも特に機微性が高い「要配慮個人情報」に該当するデータも含まれるほか、それらに該当しないまでも他人に知られたくないプライバシー性の高い情報も取り扱われることが想定され、これらの情報の漏えいや不適切な推論等が行われることで、ユーザーに深刻な影響を及ぼし得るリスクがある。

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、AI エージェント等、今後の生成 AI の発展やサービスの進展等を踏まえて随時見直すことを想定している。

2026 年 4 月に公開した本 SWG の『ヘルスケア領域における AI セーフティ評価観点ガイド』^[8]では、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、図 3-3 に示すとおりリスクの整理を行っている。

No .	評価観点	ヘルスケア領域におけるリスク概要
1	有害情報の出力制御	医療・健康に関する危険な情報（自傷・暴力の助長、根拠を欠く治療法等）が出力され、患者の生命・健康に直接的な被害をもたらすリスク
2	偽誤情報の防止	ハルシネーションにより架空エビデンスや誤った薬剤情報が生成され医療現場や生活者に影響を及ぼすリスク
3	公平性と包摂性	特定の属性の患者に対し AI の精度や品質が低下し、医療における既存の格差が拡大するリスク
4	ハイリスク利用対処	非SaMDが事実上の医療機器として利用される「目的外利用」により法規制違反等が生じるリスク
5	プライバシー保護	要配慮個人情報を含む医療データが漏えい・不正利用され、患者のプライバシーが侵害されるリスク
6	セキュリティ確保	プロンプトインジェクション等の攻撃により、医療情報の改ざんや機密データの漏えいが生じるリスク
7	説明可能性	AI 出力の根拠が不透明なまま臨床判断に用いられ、誤った医療行為や患者の不信につながるリスク
8	ロバスト性	方言・略語・非標準的な医療用語等の多様な入力に対し出力品質が不安定となり、誤った判断を招くリスク
9	データ品質	不正確または陳腐化した医療データに基づく出力が、臨床現場で誤った判断の根拠として用いられるリスク
10	検証可能性	事後検証や第三者監査が困難な状態では、問題発生時の原因究明ができず、社会的信頼を損なうリスク

図 3-3 ヘルスケア領域における AI セーフティ評価観点とリスク概要（ガイドより）

上記のリスクを踏まえ、プロダクトの開発フェーズごとに、重要な評価項目を整理し、チェックリストの形式に落とし込むことで、各事業者のガイド活用の促進を企図している。（図 3-4）

評価観点	① プロダクト設計	② モデル選定	③ プロダクト実装	④ プロダクト検証	⑤ プロダクト導入運用
有害情報の出力制御	有害情報のリスク類型を定義し、対応方針を設計	安全性ベンチマーク等で有害出力の抑制能力を評価	入力・モデル・出力の多層防御を実装	レッドチーミング・専門家レビューで抑制効果を検証	有害出力の発生状況を継続監視し、迅速に是正
偽誤情報の出力・誘導の防止	ハルシネーション等のリスクを類型化し、許容基準を定義	事実整合性・医療特化ベンチマークで正確性を評価	RAG による根拠付けと出典明示の仕組みを実装	ハルシネーション率・出典正確性を定量的に検証	ハルシネーション率を継続監視し、参照データを最新化
セキュリティ確保	LLM 固有の脅威を含むセキュリティ要件を定義	プロバイダーの体制とモデルの攻撃耐性を評価	インジェクション防御・認証・暗号化等を多層実装	ペネトレーションテスト・レッドチーミングで検証	脆弱性情報の収集・異常監視を継続的に運用

図 3-4 評価観点と 5 フェーズを掛け合わせた AI セーフティ評価（例）（ガイドより）

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とおりである。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。

2025 年度は、外部有識者へのヒアリング等によって生成 AI の利活用におけるリスク構造を明確化し、BtoB（例：医療機関向け）、BtoC（一般生活者向け）双方の代表的ユースケースを対象に、AI セーフティに関する 10 の評価観点に基づいて想定リスクの洗い出しを実施した。これらを踏まえ、サービス・プロダクトの企画・開発・運用において確認すべき評価項目を検討し、実践的なチェックリストへの落とし込みを行い、『ヘルスケア領域における AI セーフティ評価観点ガイド』として公開した。さらに、複数の具体ユースケースに対して AI セーフティの評価検証を実施し、汎用的なベンチマークでは分らない、ユースケースに特化したリスクと対応方針を確認した。

2026 年度以降、中期的には、2026 年 4 月に公開したガイドを AI 開発・提供事業者が用いて実際に現場での有用性を検証する、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のフィージビリティ検証を進める。また、ガイドの実装性を高めるため、AI 開発・提供者と AI 利用者双方の観点から、具体リスクや課題を整理し、ガイドへの反映に取り組む。

長期的な取組みとしては、技術進化や規制動向へ対応し、評価の対象・観点を随時アップデートしていくことで、リビングドキュメントとして継続的な実効性を確保する。また、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。これにより、業界全体として AI セーフティへの取組みを推進し、安全性とイノベーションの好循環を生み出すことを目指す。（図 3-5）

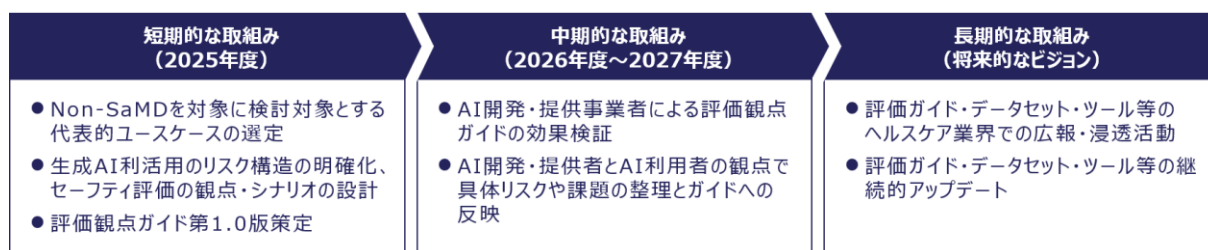


図 3-5 ヘルスケア SWG のロードマップ

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の考え方を踏まえた新たな安全性評価、すなわち AI セーフティ評価の設計と検証が喫緊の課題となっている。

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、中小企業等の組織での利用など、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。

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 (Vision Language Model: VLM) を活用した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、物理的センサ等から得られる入力に基づき、環境の認識・推論、判断・計画、又は制御出力・行動生成のいずれか一部又は全部に、機械学習・深層学習・生成 AI 等の AI 技術を用いる、自律的又は半自律的な AI システムである「Physical AI」や AI を物理システムに統合し物理世界との相互作用を可能にした「Embodied AI」等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)

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指摘されている：

- 対話型ロボット：過剰に命令的な応答や、特定の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ある。
- 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 3-6）

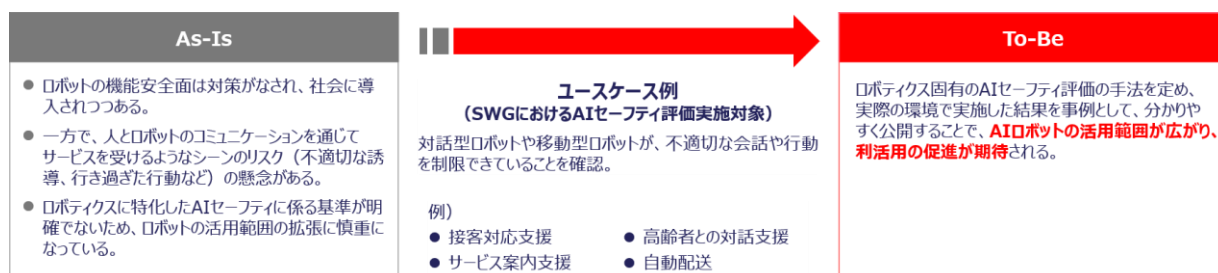


図 3-6 As-Is / ユースケース例 / To-Be の対応表

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。これらのリスクは相互に独立した排他的な関係ではなく、相互に影響し合う関係性であることを前提とする。例えば、AI ロボットが引き起こす物理的損傷を伴う事故により社会的な影響を与えることもありうる。

- 物理的リスク
ロボットの動作が人身・物的被害に直結し得るリスク（衝突、挟圧、転倒、誤動作、想定外の軌道・速度等）。
- 心理的リスク
対人応対・ケア・コミュニケーション等により、人の心理状態に影響を与えるリスク（威圧感・ストレス、依存、判断の歪み、ケアの質の偏り等）。
- 社会的リスク
サイバーセキュリティ攻撃、権利侵害や不正利用、制度未整備に起因する混乱、不公平・差別的な扱い、労働・格差等の社会的影響など、個別の事故にとどまらないリスク。



図 3-7 想定リスクの類型

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の明確化、安全開発評価プロセスの整備、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）の構築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、低減するにはどうしたら良いか」という観点を中心に、模擬環境下での評価実証及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。

2026 年 7 月に公開した本 SWG の『AI ロボティクスに関するセーフティ評価観点ガイド』^[9]では、3.2.3 で示したリスク例について評価を実施する際に必要となる評価観点を以下のとおり整理した。ただし、ここで整理した評価観点は、一部のユースケースに関連すると考えられるリスクに対するものであり、AI ロボティクス全般を網羅的に対象としたものではない。

- 物理的リスク
 - 接触、挟み込み、落下、異常動作等のリスクに対して、人に危害を加えないための「**物理的安全性**」
 - 周囲の人や環境に応じて危害を回避する「**運用合理性**」
 - 急加速・急停止等による接触事故を起こさないための「**安定性**」
 - ロボットだけではリスクを適切に低減できない場合に監督・介入を可能とする「**人間の適切な介在**」
- 心理的リスク
 - 人の意図や状態の誤認識、場違いな応答や不適切な対話、人が威圧的と感じる動作や予測しにくい振る舞い等による不安や不信等のリスクに対して、状況に即した振る舞いを確保するための「**運用合理性**」
 - 適切な振る舞いの一貫性を担保するための「**安定性**」
 - 不適切な対応が生じるおそれがある場合に監督・補正を可能とする「**人間の適切な介在**」
 - 人の意図の誤認識を避けるための「**認知・判断の妥当性**」
 - 場違いな応答や不適切な対話を防ぐための「**インタラクションの適切性**」
- 社会的リスク
 - 不公平な対応、過剰な情報収集による監視や情報漏えいの懸念、不透明な判断や責任の曖昧化等のリスクに対して、周囲への不適切な対応を防ぐ観点から監督を行うための「**人間の適切な介在**」
 - 過剰な個人情報収集や個人情報漏えいの懸念を低減する「**プライバシー保護**」

- 判断の根拠の不透明さを抑える「説明可能性」及び「検証可能性」
- 周囲への不適切な対応を抑制する「インタラクションの適切性」

また、評価方法論としては、まず AI システム単体ではなく、センサ、制御系、通信、利用環境等を含むシステム全体を対象としてリスクアセスメントを実施し、リスクシナリオを設定したうえで、設計・開発・運用の各段階で講じるリスク低減策の妥当性を分析することが重要である。評価に当たっては、シミュレーションを活用して不具合の早期発見や実機評価の効率化を図るとともに、模擬環境における評価実証（例：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 を活用した対話型ロボット実験）を通じて、実環境での運用を見据えた安全性及び有効性を確認することが求められる。さらに、必要に応じてレッドチーミング等を組み合わせ、通常の試験では顕在化しにくい誤用や想定外の入力、悪意ある利用等に対する頑健性を評価することが望ましい。加えて、各評価段階や実運用で得られた結果やインシデントを事後解析し、その結果をリスクアセスメント、評価シナリオ、シミュレーションモデル、システム設計及び運用手順の改善へ継続的に反映することが重要である。（図 3-8）

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。

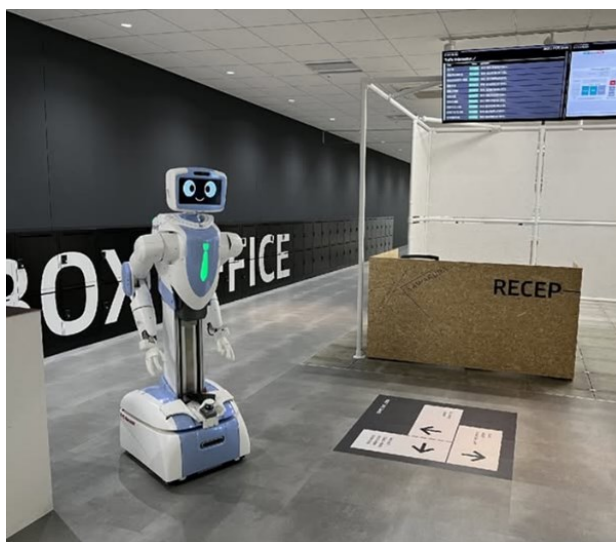


図 3-8 KAWARUBA における実証環境の概要と利用シーン例

出所)川崎重工業株式会社

3.2.5. ロードマップ

2025 年度は、人間と空間を共有し、人間とロボット、あるいはロボット同士でコミュニケーションを行う AI ロボットに関する具体的なユースケースとして、「飲食店等における商品搬送ロボット」及び「遠隔操作型の配送ロボット」を選定した。これらのロボットについて、開発・提供・運用の各フェーズにおいて生じ得るリスクを整理するとともに、導入後の運用を想定した

シナリオに基づく評価実験を実施し、その結果から抽出した AI セーフティ評価の観点を取りまとめた。

2026 年度以降、中期的には、政府の『A I ロボティクス戦略』^[10]及び『AI ロボティクス実装ロードマップ』を踏まえて、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 3-9）

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

図 3-9 ロボティクス SWG のロードマップ

3.3. 教育 SWG

2.4.2 に示したとおり、教育分野は、児童生徒や学生等の学習者、教職員向けの多様な教育分野特化の AI サービスが開発・実装され、教育課題の解決に向けた事例や実証が進められている。一方、学習過程の過度なショートカットやメタ認知的怠惰等のリスクも指摘されており、教育分野に特化したニーズやユースケースに対してリスクを踏まえた実証が実施しきれていないとの言及もある。

こうした中、教育現場における安全かつ主体的な AI 利活用の環境整備と、教育分野における適切な AI の社会実装を後押しするべく、本 SWG では、AI セーフティに関する評価観点や評価方法の明確化に取り組む。本節では、まず教育分野における AI 利活用の現状を概観しつつ、AI セーフティ評価に係る検討課題や取組み指針、AI 利活用に伴うリスク例、今後の方針（ロードマップ）を取りまとめる。

なお、教育分野には、幼児教育から高等教育やリスクリングまで様々な領域が含まれるが、本 SWG では当面は初等中等教育段階を対象として検討を行う。

3.3.1. AI 利活用の現状

近年、AI が急速に進化し社会への普及が進む中、教育分野でも AI の利活用が広がりつつある。我が国では、初等中等教育段階では、1 人 1 台端末や高速大容量の通信ネットワーク等、学校現場における ICT 環境の整備・活用を通じた教育の質向上、そして個別最適な学びと協働的な学びの一体的な充実を実現することを目指し GIGA スクール構想が推進されている。一方、少子化や教師不足・長時間勤務、学習者のニーズの多様化等への対応が学校教育を取り巻く課題として示されているところである。AI には、授業設計や教材作成支援、校務・教務の効率化、個別最適な学習への支援、特別支援や多文化多言語支援等への利活用により、学校教育を取り巻く課題解消の一助となることが期待されている。

こうした中、文部科学省は、2023 年 7 月に『初等中等教育段階における生成 AI の利用に関する暫定的なガイドライン』を公表し、2024 年 12 月にガイドライン第 2.0 版^[11]を公表している。その後も、生成 AI パイロット校事業や教育分野特化の生成 AI の実証事業等を通じ、教育現場での生成 AI の活用可能性やリスクに関する知見の蓄積を進めている。

国際的にも、AI 利活用を通じた児童生徒のパーソナライズされた支援、校務・教務の効率化、教育システム・教育機関の運営支援に関する研究が進んでいる。例えば、『OECD Digital Education Outlook 2026』^[12]は、教育分野における AI の効果が教育的意図や利用設計に左右されること、課題の遂行や成果物の質が向上しても、必ずしも学習成果や知識の定着を意味しないことを指摘している。

こうした指摘に加え、教育分野における AI 利活用をめぐっては、誤情報、バイアス、不適切な出力等の一般的なリスクと併せて、学習過程の過度なショートカットや思考の依存、メタ認知的怠惰による学習過程への影響等の論点が議論されている。

我が国においては、これらの可能性とリスクを踏まえ、初等中等教育段階における AI の支援による教務・校務の負担軽減・教師の専門性の向上、多様な学習ニーズへの対応や包摂性の向上に資する AI の実装、さらには、深い学びを通じた児童生徒の資質・能力の育成に資する AI の活用・実装が検討されている。その際には、発達の段階や学習過程に及ぼす影響等のリスクを適切に考慮することが求められる。

3.3.2. AI セーフティ評価の重要性と利活用促進への期待

上述のとおり、初等中等教育段階においては、児童生徒向けの学習支援や教師向けの授業・校務支援等に生成 AI を活用する取組みが進められている。こうしたプロダクト・サービスについては、精度や利便性のみならず、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となる。特に、生成 AI の出力や利用過程が児童生徒の学習・発達や教師の判断に及ぼす影響を、利用場面に即して検証できる枠組みの整備が課題であり、AI セーフティ評価の重要性が高まっているといえる。

文部科学省の『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』では、学校現場で生成 AI を利活用する際の基本的な考え方や留意事項が示されている。また、安全かつ主体的に AI を利活用できる学習環境の構築に向けた取組みとして、教育現場のリスク整理、評価基準及び評価データの整備が検討されている。一方、個別の学習場面等のユースケースにおける評価シナリオ・評価項目及び評価方法にはなお整理の余地があり、今後整理が進み、広く関係者に示されることで、AI セーフティが確保され、初等中等教育段階での適切な生成 AI の利活用が促進されることが期待される。

3.3.3. 想定されるリスク例

公的資料・関連研究を踏まえ、教育分野において以下のようなリスク例が想定されている。今後、ユースケースに基づいて本 SWG にてリスク例の議論・検討を進め、AI セーフティに関する評価観点ガイドに記されている評価観点を基に整理・精緻化を目指す予定である。

- **AI への情緒的依存**

児童生徒が AI を擬人化する傾向が、発達途上にある認知的・社会情動的能力と相まって、チャットボットとの相互作用における依存、情緒的愛着及び搾取を生じさせるリスク

- **生産的な試行錯誤の喪失**

生成 AI への過度な依存が、学習方略を調整するメタ認知的な関与を弱める「メタ認知的怠惰」を招き、生産的な試行錯誤（productive struggle）を失わせるとともに、学習者のエージェンシーを低下させるリスク

- **AI 依存による成績低下**

AI の支援を受けて課題を遂行できたことを、児童生徒が自らの理解や能力によるものと

誤認し、実際の習熟度を過大評価する可能性がある。その結果、AI を活用した後に AI 支援なしでテストをした場合、成績が低下するリスク

- **プロンプトインジェクションによる評価操作**

攻撃者として行動する児童生徒が、一見通常の課題提出物に悪意ある指示を埋め込み、モデルの挙動に影響を与えて、より有利な評価につながるリスク

3.3.4. AI セーフティ評価の方向性

本 SWG では目下、AI 提供者（EdTech 事業者等）が提供する AI プロダクト・サービスを対象に、初等中等教育段階において児童生徒が生成 AI を利活用する際に生じ得るリスクを検討する。具体的には、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など、様々な相談対応等の場面が想定される。なお、児童生徒への影響は、教師による生成 AI 利活用とも関連すると考えられるため、教師による教材作成や児童生徒の状況把握・評価等についても検討の対象として含める。

教育分野では、同一のモデルや出力でも、児童生徒の発達の段階、学習目的やその内容等に応じて、影響の大きさや許容されるリスクが異なる場合がある。このため、基盤モデル（LLM）だけを切り離して評価するのではなく、ユースケースごとに、誰が、何の目的で、どのように利用するか等を明確にし、影響を受ける主体、望ましくない事象及びその発生要因を整理してリスクシナリオを具体化する。それぞれのリスクに対する既存または想定されるガードレール等のリスク低減策を整理した上で、有効性を確認するためのセーフティ評価観点・項目を検討し、ガイドやチェックリストの作成も視野に検討を進めていくことを想定している。

3.3.5. ロードマップ

本 SWG では本年度（2026 年度）を活動初年度として、教育における主要なユースケース及びリスクを体系的に整理することを重視し、多様な観点を取り込むべく、アカデミアや事業者、学校現場や関係省庁など各所に対してヒアリング等での意見交換を実施する。その上で、整理したユースケース及びリスクを AI セーフティ評価観点に紐づけつつ、各リスクの評価観点・項目をガイド形式で取りまとめていく。

また今後は、AISI の役割を踏まえ、AI セーフティの評価指標の策定・評価に資するべく、初等中等教育領域における評価観点ガイド、AI の安全性ベンチマーク及び評価環境の整備も見据えていく。

短期的な取組み (2026年度)	中長期的な取組み (2027年度～、将来的なビジョン)
● 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	● 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

図 3-10 教育 SWG のロードマップ

3.4. ジオ AI SWG

地理空間情報（Geospatial Information）は、人流・物流、インフラ、環境、防災、都市計画、安全保障など幅広い領域を支える基盤データとして位置付けられており、『人工知能基本計画』においても AI 利活用の加速的推進の具体的な施策の一つとして「ジオ AI」が取り上げられるなど、AI との融合による地理空間情報の価値の飛躍的な向上に対する期待が高まっている。一方で、地理空間情報と AI の高度な融合・活用は、プライバシー侵害や監視の強化、誤情報の拡散、さらには重要インフラに対するサイバー攻撃の高度化など、新たなリスクをもたらす可能性を有している。また、航空写真、衛星データ、3D 都市モデル、リアルタイムセンシングなど、多様なデータソースの統合が進展する中で、データの信頼性や品質の確保に加え、AI システムの透明性や説明可能性をいかに担保するかが重要な課題となっている。

本節では、AISI の共通的な AI セーフティに関する評価観点を踏まえ、地理空間情報の品質・信頼性に加え、プライバシー、セキュリティ、不適切な推論・利用、説明可能性等のジオ AI に係るリスクをユースケースに即して整理する。あわせて、具体的な評価フレームワーク及びガイドラインの検討の方向性を示すとともに、国際標準化機関や国際的な政策対話の場への提案を見据えたロードマップを記述する。

3.4.1. AI 利活用の現状

内閣官房地理空間情報活用推進室及び国土交通省が立ち上げたジオ AI 研究会が公表した『中間整理～地理空間情報×AI の融合の推進～』^[13]では、AI 技術の急速な進展を背景として、地理空間情報分野においてもデータ整備、可視化・データ分析などの機能で AI 活用が既に実用段階にあることが示されている。

地理空間情報と AI の融合には、大きく分けて二つの方向性が存在する。一つは、「AI for 地理空間情報」（ジオ AI 研究会『中間整理』では「AI for GIS」と表記）であり、地理空間情報の整備、更新、可視化などのプロセスに AI を活用する方向性である。もう一つは、「地理空間情報 for AI」（同「GIS for AI」）であり、様々なユースケースにおいて、他のデータと組み合わせながら地理空間情報を AI による解析やシミュレーションに活用する方向性である。

前者では、これまで主として人手により整備・作成されてきた地理空間情報（例えば、地形図や 3D 都市モデル）について、AI を活用した自動生成や更新の高度化が進展している。後者では、浸水リスク予測やインフラ劣化診断、都市計画等において、地理空間情報を他のデータと組み合わせ、AI による分析・予測や意思決定支援に活用する取組みが進められている。

3.4.2. AI セーフティ評価の重要性と利活用促進への期待

地理空間情報は、公共測量、防災、インフラ管理、都市計画等、社会・経済活動を支える様々な用途で利用されているため、AI による地理空間情報の整備・更新や分析・推論に誤りが生じた場合、その影響が後続する分析や意思決定に波及する可能性がある。詳細な位置情報等を取り扱うため、プライバシーやセキュリティへの配慮も重要となる。

このような地理空間情報の特性を踏まえ、AI システムの性能・信頼性のみならず、生成・更新された成果データの品質や、利用文脈に応じたリスクを評価できる枠組みを整備することが重要である。評価方法や基準が明確になることで、AI 開発・提供事業者や地理空間情報の整備・利用者が、AI システム及び成果の品質・信頼性を客観的に確認・説明できるようになり、ジオ AI の適切な導入・利活用の促進につながることが期待される。

3.4.3. 想定されるリスク例

地理空間情報と AI の融合を社会実装の観点から捉えると、以下のような課題が存在する。これらの課題は、AI を活用した地理空間情報の整備・更新や、地理空間情報を用いた AI による分析・推論等において様々な問題につながる可能性がある。本 SWG では、こうした課題とそれに起因するリスクを整理していく。

AI for 地理空間情報の観点

- 公共領域における地理空間情報の整備（公共測量）については、製品仕様、整備プロセス、品質管理及び品質評価手法が関係法令や規程によって定められている。しかし、これらの枠組みは AI 技術の活用を前提として設計されたものではなく、現行制度上の位置付けや品質評価の考え方を整理することが、AI を活用した整備・更新手法の導入に向けた課題となっており、必要な品質を満たしていない地理空間情報が利用されるリスクが考えられる。
- 地理空間情報の整備・更新に活用される AI システムについては、その性能や信頼性を客観的に評価するための基準や評価手法が十分に確立されていない。また、ベンチマークデータセット、評価指標、評価結果の開示・表示方法等に関する共通ルールも未整備であることから、AI 技術の導入効果や信頼性を第三者に対して説明・証明することが困難であり、利用目的に対して十分な信頼性を有しない AI システムが用いられるリスクがある。
- AI を活用して整備・更新された地理空間情報の成果については、品質評価の手順や評価手法が十分に体系化されていない。そのため、成果データの位置正確度、完全性、一貫性及び信頼性を客観的に評価し、利用者に対して品質を説明するための仕組みが不足しており、分析や判断に影響を及ぼすリスクが考えられる。
- AI を活用して浸水リスク予測やインフラ劣化診断などの地理空間情報データ分析・可視化などを行った結果に対する信頼性や誤率の表示方法等に関する共通ルールが未整備であることから、説明責任や透明性が不足しており、不適切な意思決定につながるリスクがある。

地理空間情報 for AI の観点

- AI が現実空間における対象間の関係性を適切に理解し、データの取得・選択から解析、推論、意思決定支援までを一貫して実行するためには、地理空間情報特有の位相関係、空間参照系、空間演算等を適切に扱うことのできる AI システム、すなわち地理空間基盤

モデル（geospatial foundation model）の実現が求められる。現状、こうした地理空間情報特有の構造や意味を AI が十分に理解・処理するための技術的枠組みは、依然として発展途上にあり、例えば位置関係等が誤って解釈され、不適切な分析・推論につながるリスクがある。

- 地理空間基盤モデルの安全性、信頼性及び頑健性を評価するための評価手法や評価指標は、現時点において十分に整備されていない。特に、安全保障上の問題やプライバシー侵害の可能性、不適切な推論や情報利用の誘発といったリスクがあり、それらをどのように検出・評価するかが明確になっておらず、多様なユースケースや利用環境を踏まえた体系的な評価・検証の枠組みの構築が求められている。

3.4.4. AI セーフティ評価の方向性

本 SWG では、ジオ AI における AI セーフティ評価の具体化に向け、①地理空間情報を活用した AI システムにおけるリスク評価及び安全確保のあり方に関すること、②地理空間情報×AI ユースケースにおける AI リスク分析に関すること、③国際的な議論動向を踏まえた標準化及びルール形成への貢献に関すること、④その他、社会・業界動向を踏まえた検討課題に関すること、の4つのテーマを柱として取り組む（表 3-2 参照）。①及び②を AI セーフティ評価の具体化に向けた取組みとし、③ではその成果の標準化・ルール形成への展開、④では技術・社会環境の変化に伴って生じる新たな課題への対応を検討する。特に、ルール策定が急がれる地理空間情報整備・更新、AI システムの評価、及び整備成果の品質評価（検討テーマ①のデータ整備に該当）については早期に取り組む。

表 3-2 ジオ AI SWG の検討テーマ及び取組み方針

検討テーマ	機能			
	データ整備		可視化・データ分析	予測・提案・行動
	a. データ取得、整備・加工	b. データ管理、保管	c. API提供、オープンデータ化・データ販売等	d. 意思決定支援、サービス開発等
① 地理空間情報を活用したAIシステムにおけるリスク評価および安全確保のあり方に関すること	◎ (AIシステム、品質評価)	◎ (メタデータ、セキュリティ、第三者認証)	—	—
② 地理空間情報×AI ユースケースにおけるAI リスク分析に関すること	○	○	○	○
③ 国際的な議論動向を踏まえた標準化及びルール形成への貢献に関すること	◎ (品質評価等)	◎ (メタデータ等)	○	○
④ 社会・業界動向を踏まえた検討課題に関すること	○	○	○	○

3.4.5. ロードマップ

2026 年度は、地理空間情報の整備に従事する産業界関係者を中心として SWG を組成し、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を検討する。また、具体的な事例として 2 次元マップ（建物）を対象としてガイドラインを検討する。

2027 年度以降、中期的には、2 次元マップ（建物）以外の地理空間情報を対象としてガイドラインの範囲を徐々に拡張していく。また、地理空間情報×AI ユースケースにおけるリスク分析及びガイドライン検討を行う。2026 年度からの成果も含め、国際標準化団体である ISO/TC 211 及び OGC (Open Geospatial Consortium) に提案を行う。

将来的には、国際標準化への貢献を進めるとともに、AI システム及び成果データの評価結果の活用や、認証を含む適合性評価のあり方について、分野横断 SWG や関係機関との連携も踏まえて検討を進める。（図 3-11）

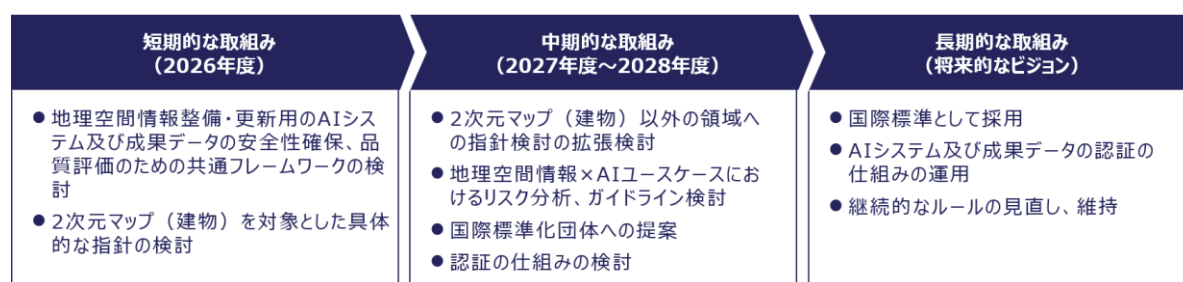


図 3-11 ジオ AI SWG のロードマップ

4.

分野横断 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、分野横断 SWG で検討されている取組みの全体像を整理する。データ品質 SWG 及び適合性評価 SWG では、分野別・具体的ユースケースごとの検討を相互補完的に支えるため、データ品質や適合性評価の観点から、AI セーフティ評価の枠組みを実務に適用していくための共通的な考え方を整理する。

4.1. 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利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。特に、データ品質を共通の基準で評価・比較し、各組織における継続的な品質改善や、品質が不十分なデータに起因する AI の誤作動・事故の防止に活用できる共通の評価尺度を整備することを目指す。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。

4.1.1. 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に大きく左右される。これは AI セーフティの前提条件であり、AISI の『AI セーフティに関する評価観点ガイド』においても、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った領域であった。生成 AI の普及に伴い、データ品質への関心は AI の開発・提供事業者から利用者にも広がり、さらにマルチモダリティの発展、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抑制する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。

このような背景から、従来の人間による利用や、基幹系システム、BI 等での処理を主眼として整備されてきたデータに加え、AI が安全かつ有効に利用できる高品質な AI-Ready Data として整備することの重要性が高まっている。

AI-Ready Data の整備に当たっては、正確性、完全性、一貫性、最新性等の観点からデータそのものの品質を高めることに加え、AI がデータの意味と適切な利用方法を判断できるよう、データの意味、来歴、利用条件、責任主体、想定される用途や適用範囲等の利用文脈を、機械可読性の高いメタデータとして付与・管理することも含めて、データ品質管理として捉える必要がある。

さらに、Quality by Design の考え方にに基づきデータの設計段階から品質要件や評価方法を組み込むとともに、運用中はデータやメタデータの状態を機械的・継続的に評価し、品質劣化の兆候や変化を捉え、是正につなげる仕組みが求められる。あわせて、重要な判断については Human-in-the-Loop により人が確認・承認し、その根拠と責任分界を記録することで、データ品質の検証可能性・説明可能性を確保することが重要となる。

現在、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。AI システムやその開発・運用に用いられるデータは、国境を越えて流通し、広く開発・提供・利用されるものである。このため、国や組織ごとに異なる基準だけでは、データ品質に関する共通理解や評価結果の比較可能性を確保することが難しく、国際的に共有可能な評価枠組みが重要となる。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』^[4]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。

- 網羅的なデータ品質の取組み検討の難しさ

データ品質は単一の概念ではなく、正確性・完全性・一貫性など複数の品質特性を含む。また、技術・組織・ライフサイクル全体の運用など複数のレイヤーでの取組みが求められる。このため、どの取組みを優先すべきか判断することが難しく、全体像を整理しないまま、特定のリスクや技術的課題への対応に偏った部分最適的な施策に終始して

しまうケースが生じやすい。こうした抜け漏れを防ぐため、データ品質に関する取組み状況を簡易的かつ網羅的に確認できる枠組みやツールの必要性が指摘されている。

- データ品質のユースケース依存性

データ品質において重視すべき特性は、ユースケースの目的や利用するデータの特性等によって異なる。例えば、AI モデルの学習ではラベルの正確性や偏りの少なさが重要となる一方、経営ダッシュボードでは集計値の整合性が重要となる。この違いを整理しないまま一律の基準を適用すると、過剰品質によるコスト増大や、逆に品質不足によるリスク拡大につながりかねない。したがって、ユースケース起点で重要となる品質特性を定義し、必要な取組みを選択するアプローチが求められる。

- 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付けにあるため、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。

4.1.2. 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』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 4-1）。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及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。

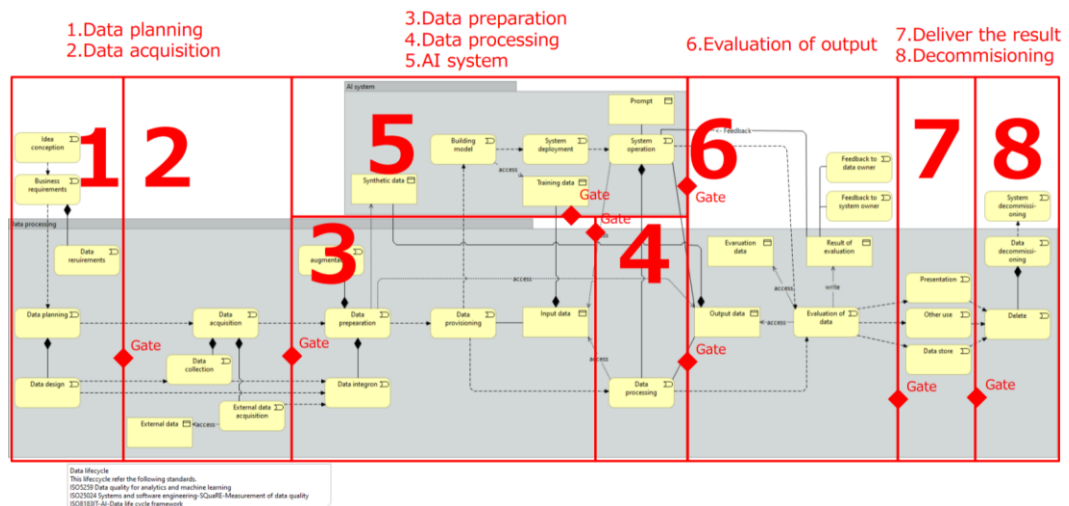


図 4-1 AI システムも含めたデータライフサイクル

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。（図 4-2）

その中核となるのが、データ品質を共通の基準で評価・比較できる「評価尺度（メジャー）」の整備である。共通のメジャーを確立することで、各組織が自らのデータ品質の状態や課題を把握し、継続的な改善につなげることが可能となる。また、AI システムの目的やリスクに応じて求められる品質水準の違いを考慮し、利用するデータが当該ユースケースに対して十分な品質を備えているかを判断できるようにすることも、評価尺度の重要な役割である。これにより、品質が不十分なデータの利用に起因する誤出力や事故のリスクを低減し、安全な AI の開発・提供・利用を支えることを目指す。

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、ユースケースやリスク特性を踏まえながら、現場における適用検証を通じて実証的に検討を進める。



図 4-2 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）など。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。評価結果及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。

4.1.3. ロードマップ

2025 年度は、データ品質管理の高度化に向けた課題解決に資するため、ユースケースの特性に応じて必要なデータ品質の取組み整理、データ品質マネジメントチェックリストの作成・検証、官報データを題材として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を実施した。

2026 年度以降、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 4-3）

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)
● 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）	● 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	● 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

図 4-3 データ品質 SWG のロードマップ

4.2. 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り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、ISO 9001 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 を基準とした第三者適合性評価制度の整備も進んでいる。

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 の国際標準化も進行中である。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）及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。

4.2.1. 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で行う必要がある。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従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。

4.2.2. 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、『AI セーフティに関する評価観点ガイド』を踏まえて開発される評価ツール等を使った分野別 SWG の具体的な取組みから、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。

具体的には、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での取組みへの反映を図る。

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 4-4）

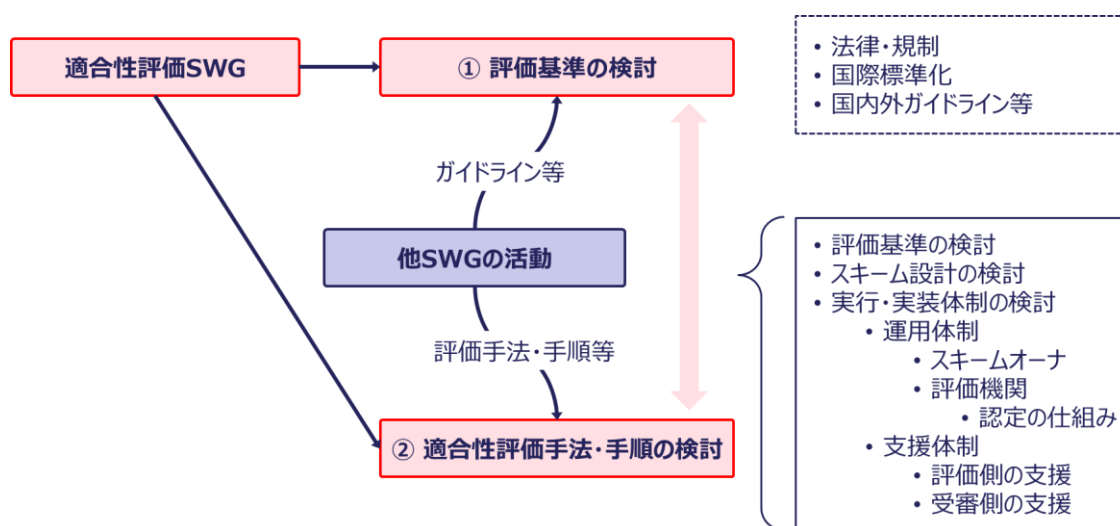


図 4-4 適合性評価手法の確立への取組み方針

4.2.3. ロードマップ

2025 年度は、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、ISO/IEC JTC1 SC42 にて規格開発中の Human-Machine Teaming (HMT) を題材として、包括的な AI 適合性評価の論点を整理し、評価観点や課題、今後の制度設計に向けた方向性について検討を行った。また、分野別 SWG と連携し、HMT や評価ニーズに関するヒアリングを実施した。

2026 年度以降、中期的には、評価側の体制を拡充するとともに、受審側組織との連携も強化する。あわせて、ISO/IEC 42007 を含む ISO/IEC JTC1 SC42 における国際標準化の議論を踏まえて、包括的なアシュアランス及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取り組む。

その中核となる考え方が、図 4-5 に示す CCAS（Compositional Conformity Assessment Schemata）である。CCAS では、組織内の各階層における多層的な評価にとどまらず、サプライチェーンやバリューチェーン等を通じた組織間の関係、すなわちトラストチェーンも評価対象に含める。各階層の評価結果を、組織間の関係性や責任分担と併せて総合的に確認することにより、AI セーフティの達成状況を包括的に評価できる仕組みの構築を目指す。

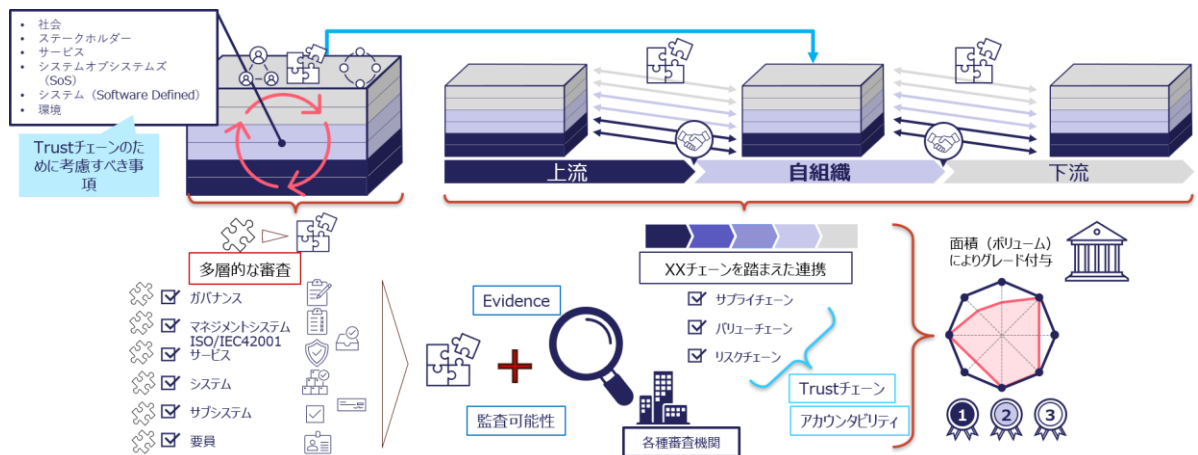


図 4-5 CCAS（Compositional Conformity Assessment Schemata）

今後は、この考え方にに基づき、相互運用や相互承認も視野に入れた実現可能な制度設計を検討するとともに、国内における AI に関する適合性評価制度の具体化と試行に着手し、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 4-6）

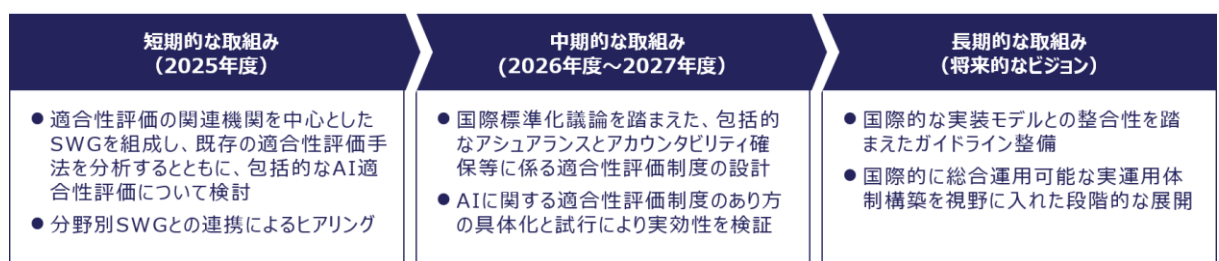


図 4-6 適合性評価 SWG のロードマップ

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を構築することにより、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。

- [1] 総務省／経済産業省『AI 事業者ガイドライン』（2026 年 3 月 第 1.2 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/20260331_report.html
- [2] AISI『AI セーフティに関する評価観点ガイド』（2026 年 7 月 第 1.20 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ムing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 AI メンタルヘルスケア協会『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）』（2025 年 4 月 公表）
<https://www.aimh.or.jp/guideline>
- [6] 『ロボット産業ビジョン 2050』Ver.1.0（2025 年 6 月 公表）
<https://www.jara.jp/publications/visionver1.html>
- [7] 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』（2025 年 7 月 第 1.1 版公表）
<https://www.fdua.org/news/20250714-02>
- [8] AISI『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 第 1.0 版公表）
https://aisi.go.jp/output/output_information/260402/
- [9] AISI『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 第 1.0 版公表）
https://aisi.go.jp/output/output_information/260723/
- [10] A I ロボティクスに関する関係府省連絡会議『A I ロボティクス戦略 ～社会実装を加速し、巨大市場を切り拓く～』（2026 年 3 月 公表）
https://www.cas.go.jp/jp/seisaku/ai_robo/dai2/shiryo1.pdf

- [11] 文部科学省『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』（2024 年 12 月 Ver.2.0 公表）
https://www.mext.go.jp/a_menu/other/mext_02412.html
- [12] Organization for Economic Co-operation and Development『OECD Digital Education Outlook 2026』（2026 年 1 月 公表）
https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/01/oecd-digital-education-outlook-2026_940e0dd8/062a7394-en.pdf
- [13] ジオ AI 研究会『中間整理 ～地理空間情報×AI の融合の推進～』（2026 年 5 月 公表）
<https://www.gsi.go.jp/chirikukan/chirikukan61037.html>
- [14] AISI『データ品質マネジメントガイドブック』（2026 年 5 月 Ver.1.02 公表）
https://aisi.go.jp/output/output_framework/data_quality_management_guidebook/

更新履歴

Ver	更新日付	主な更新内容
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
2.0	令和 8 年（2026 年）10 月	既存 SWG のビジョンを更新 新設 SWG のビジョンを追記