AI ロボティクスに関するセーフティ評価観点ガイド

令和 8 年 7 月 23 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

ロボティクスサブワーキンググループ

目次

1. 背景・目的

1.1. 背景

近年、ロボティクス分野において、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・制御が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、従来は定型作業を前提とした設計が主流であり、対象物の不定形性や環境変動への対応に対する難度や導入コスト等の課題から、食品・医薬品等の産業分野、物流搬送等の用途、ならびに中小企業等での利用は限定的であった。しかし近年では、多品種少量の利用領域（いわゆるロングテール領域）での活用が拡大しつつあり、こうした領域では従来よりも柔軟かつ容易にロボットが利用できることが求められるため、AI の活用により汎用性を増した協働ロボットへの期待が高まっている。

このようなニーズの高まりと技術の進展により、AI を活用した自律移動型ロボットや対話型ロボット、協働ロボット等の多種多様なロボットが今後ますます普及し、認知・判断・行動の各段階における AI の関与が拡大していくと考えられる。一方で、AI 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、AI による認識や学習等に起因して生じ得るリスクに対して、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と文脈理解、人に対する支援といった観点が追加的に必要となる中で、ロボティクスにおける従来の機能安全を前提とし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である。

1.2. 目的

本書は、上記のような背景と課題認識を踏まえ、AI の利用に伴って新たに生じ得るリスクを特定・評価し、必要に応じたリスク低減策を検討するための観点を示すことにより、事業者が AI セーフティに配慮した AI ロボットのサービス・プロダクト開発・設計・提供・運用を行うことができるようにすることを目的とする。

具体的には、AI ロボティクスにおける特有のリスクを考慮した AI セーフティを評価するための観点を示す。これにより、AI ロボット導入時に考慮すべき安全性・運用上の観点を整理し、AI ロボティクス事業者による AI ロボットの安全性向上に配慮した社会実装と、持続的なビジネス価値創出の両立を図る。

また本書は、産・学・官の連携を前提としたリビングドキュメントとして、生成 AI を含む AI

技術や社会の変化を反映しながら適宜改版することを予定している。

1.3. 本書の位置づけ

1.3.1. 本書のスコープ

本書は、AI を搭載し、人と空間を共有しながら移動、対話、搬送等を行う AI ロボットを対象とする。実機を使って行ったセーフティ検討の過程で得られた知見を踏まえ、開発・提供・運用の各フェーズにおいて生じ得るリスク及びその評価観点を提示する。

対象とする AI は、テキストベースの大規模言語モデル（Large Language Model： LLM)、画像認識 AI、音声認識 AI とし、これらを用いて実現する様々な認識、判断、対話、指示理解、移動経路生成等の各機能を、対象とする AI 機能とする。

なお本書においては、上記のとおり対象とする AI を限定した形としたが、次版以降、Vision Language Action（VLA）やロボット基盤モデル等、様々な AI ロボティクスを検討対象とする。

1.3.2. 想定するユースケース

上記のスコープに基づき本書では、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を想定している。その内容を表 1-1 に示す。本書の評価観点は、これらのユースケースに関するセーフティ評価実験から抽出したものである。

表 1-1　本書で想定するユースケースと主な評価観点

ロボットの種類	ユースケースと主な評価観点
飲食店等における 商品搬送ロボット	・飲食物等を利用者に搬送するロボットが、音声による注文受付等のコミュニケーションを行うとともに、店舗内を自律移動しながらサービス提供を行うユースケースである。 ・本ユースケースでは、屋内環境における移動、人とのすれ違い、障害物回避、転倒・衝突防止等が安全を考える上で重要となる。また、子供の不用意な接近等、利用者との接触可能性を考慮した安全設計や、注文時等の会話で、利用者に不安や威圧感を与えないことも求められる。 ・さらに、AI による経路生成や環境認識の誤り、センサ誤認識等に起因する誤動作への対応、異常発生時の停止・人間の介入手段の確保等も重要な評価観点となる。
遠隔操作型の 配送ロボット	・オペレータが、複数の小型配送ロボットを遠隔で監視し、必要に応じて介入することも想定したうえで、実験施設内や公道で荷物の配送を行うユースケースである。 ・当該ユースケースでは、通信遅延・通信断に起因する情報不足への対応、遠隔オペレータの状況認識支援、操作権限管理等が重要とな

	る。また、周辺歩行者や他車両との相互作用を踏まえた安全確保、危険を察知した場合の安全な停止や自律退避等も重要な論点となる。 ・遠隔操作と AI による自律機能が組み合わさることでオペレータの介入タイミングが複雑化するため、多様なステークホルダーの責任分担や、AI の判断の透明性、各種ログの取得、説明可能性の確保等も重要な評価観点となる。

1.3.3. 対象読者

2026 年 3 月に AI ロボティクスに関する関係府省連絡会議により公表された「AI 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～」[1]では、AI ロボティクスの需要側と供給側の基本的な考え方や取り組みが示されている。本書では、主に供給側を対象とし、その中で、特に AI ロボティクスの開発や提供・運用に直接携わるステークホルダーを主な読者としている。具体的には、総務省・経済産業省の「AI 事業者ガイドライン」[2]における「AI 開発者」及び「AI 提供者」の AI 側のステークホルダーである。

なお、ロボットメーカー、ロボット SIer、コンポーネントメーカー等は、その事業内容や提供形態に応じて、AI 事業者ガイドライン上の AI 開発者又は AI 提供者に該当し得るため、本書では、これらの主体を厳密に一対一対応させるものではなく、AI ロボティクスの開発・提供・運用に関与する供給側の主体が、自らの役割に応じて第 3 章及び第 4 章の観点を参照することを想定する。

また上記のとおり、主に供給側を対象としているため、AI 事業者ガイドラインにおける「AI 利用者」は直接的には対象としていないが、AI ロボティクスを利用するうえで、本書で示す AI セーフティに関する考え方はぜひ参考にしていただきたい。

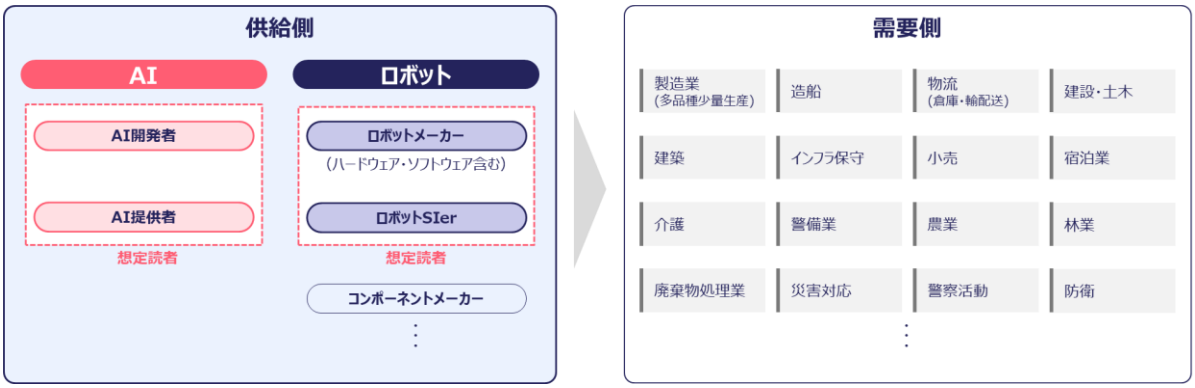


図 1-1　AI ロボティクスのステークホルダー（イメージ）

1.3.4. 本書及び別冊の使い方

本書は、AI ロボティクスの社会実装に向けて、AI セーフティに関する背景の説明から具体的な評価観点の提示、将来課題の整理に至るまでを段階的に構成している。各章の位置づけと目的は以下のとおりである。また全体の構成イメージを図 1-2 に示す。

第 2 章「AI 利活用・セーフティの動向」

AI ロボティクスを取り巻く技術動向、市場動向、政策・規制動向を示す。国内外の制度的枠組

[1] ＡＩロボティクスに関する関係府省連絡会議「ＡＩ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）https://www.cas.go.jp/jp/seisaku/ai_robo/dai2/shiryo1.pdf

[2] 総務省・経済産業省「AI 事業者ガイドライン（1.2 版）」（2026 年 3 月）https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20260331_report.html

みや標準化動向を整理し、本ガイドが、国際的な AI セーフティの議論と整合する位置づけにあることを示す。

第 3 章「AI ロボットのリスクと要因」

評価に先立ち、AI ロボティクスにおける特有のリスクを分類する。危害の対象として「物理的リスク」「心理的リスク」「社会的リスク」の三類型を整理し、それらが相互に影響し合う構造を示すとともに、「AI」「ロボット」「人間」「社会」の四要因による分析枠組みを提示する。

第 4 章「AI ロボティクスに関するセーフティ評価観点」

第 3 章で整理したリスクを評価するための評価観点について、それらを実用的に活用できるように整理する。従来の AI セーフティやロボットの機械・機能安全から、AI とロボットが融合することで生じる新たな評価観点を対象としてまとめる。

第 5 章「今後の課題と方向性」

AI ロボティクスが今後どのように発展・社会実装に向けて進むかを踏まえ、今後の課題や方向性を示す。

また、本書の他に「実証実験報告書」及び「評価手法解説書」を別冊として用意する。本書及びこれら別冊を一体的に参照することで、AI ロボティクスに関する安全性への理解が促進されることを期待する。

別冊「実証実験報告書」

1.3.2 項で示したユースケースについて実施した実証実験の目的、本書で示す評価観点に基づく評価シナリオ、評価方法、評価結果及び考察を示した報告書である。

別冊「評価手法解説書」

AI ロボティクスに関する安全性評価において、重要となる評価手法を概説した解説書である。実証実験における活用方法を、具体事例を含めて示す。

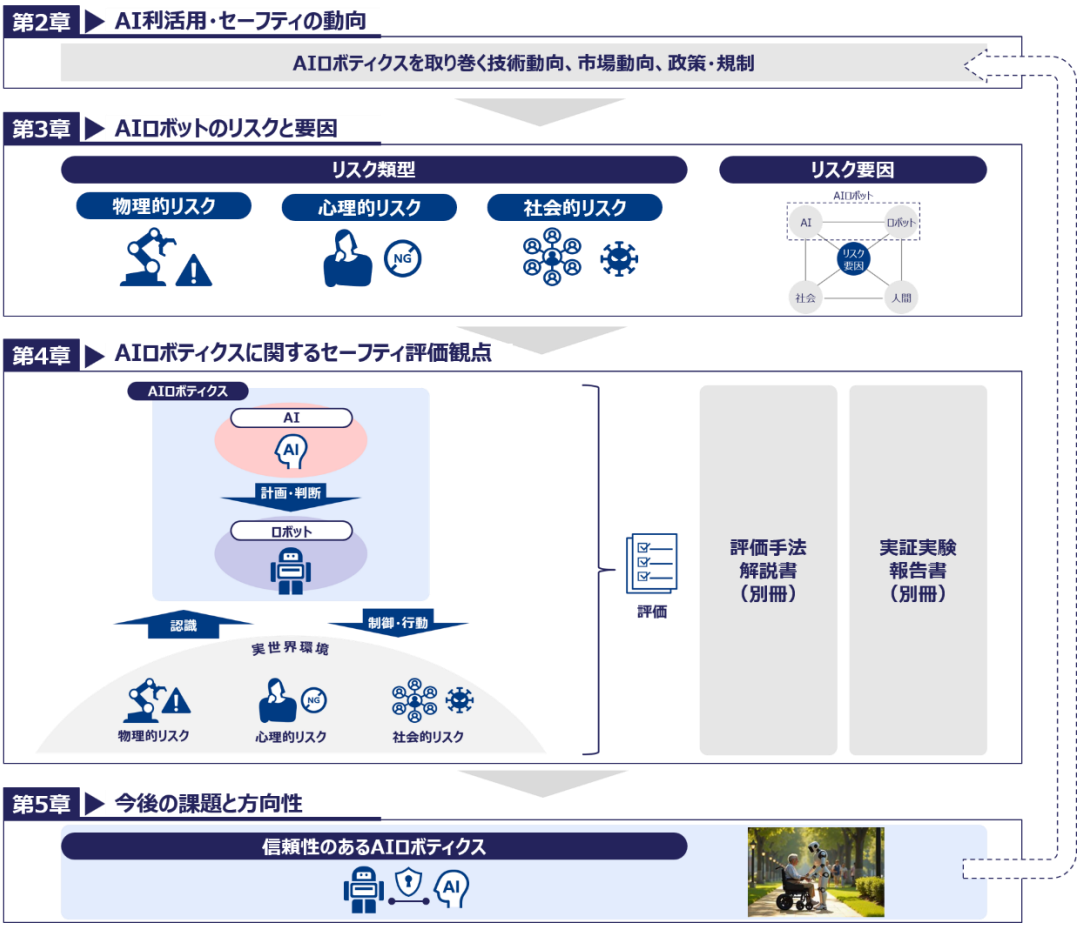
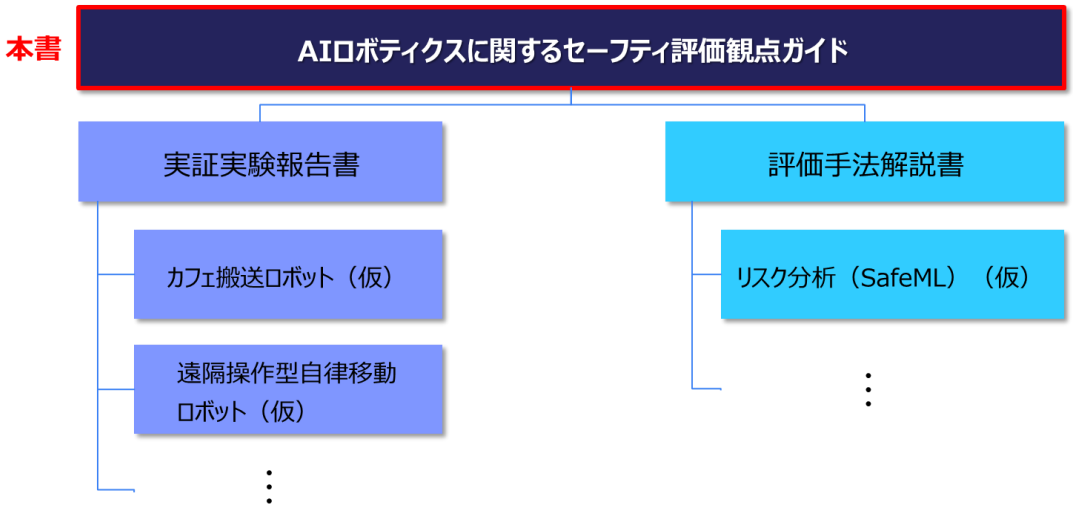


図 1-2　本ガイドの構成イメージ

本書と別冊の「実証実験報告書」及び「評価手法解説書」の関係性を図 1-3 に示す。「実証実験報告書」は、実証実験のユースケースごとに報告書を作成していくことを想定している。また、「評価手法解説書」は、今後新たに用いる評価手法ごとに解説書を作成していくことを想定している[3]。



[3] 別冊として、「カフェ搬送ロボット（仮）」、「遠隔操作型自律移動ロボット（仮）」の実証実験報告書及び「リスク分析（SafeML）（仮）」の評価手法解説書を今後公表予定。

図 1-3　本書を含む文書の構成イメージ

AI ロボティクスの開発、提供、運用の実務で本書を利用する際は、第 3 章及び第 4 章を中心としつつ、必要に応じて第 2 章を参照することを推奨する。また、以下に代表的な場合に対して、特に参考にすべき章を整理する。

・AI ロボティクスの新規導入を検討する場合：第 1 章、第 2 章、第 3 章、第 4 章
・AI ロボティクスの設計・開発・提供・運用を行う場合：第 3 章、第 4 章
・事故・トラブル発生後に見直す場合：第 3.2 節、第 4 章
・技術動向、利活用事例、政策動向を把握したい場合：第 2 章

なお、本書は、AI ロボティクスに関する一般的なセーフティ評価観点を整理したものであり、特定用途・特定環境における安全性、法令適合性、性能又は品質を保証するものではない。

実際の導入・運用にあたっては、個別用途、利用環境、対象利用者、適用法令、関連規格、契約条件等を踏まえ、事業者が独自にリスク評価及び安全確認を実施する必要がある。

1.4. 本書で使用する用語

使用する用語の本書における定義を以下に示す。

リスク

本書では、AI ロボットの利用により、利用者、周囲の人、組織又は社会に望ましくない影響が生じる可能性及びその影響の大きさをいう。具体的には、物理的リスク、心理的リスク及び社会的リスクを中心に扱う。

AI ロボティクス

物理空間で動作するロボットの知覚、計画、判断、制御等の機能と AI を統合して用いる技術であり、LLM 基盤モデルやロボット基盤モデル等の活用を前提に、柔軟な指示理解や高度な Human-Machine Teaming（HMT）などを実現可能なものを指す。

AI ロボット

センサ、駆動系、制御系の 3 つの要素技術を有し、LLM、画像認識、音声認識等の AI を搭載して知能化された機械システム。

AI セーフティ

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[4]

End to End

本書では、AI ロボティクスの文脈において、環境認識、判断、動作生成・制御等の複数モジュールで構成される処理を統合し、入力から出力までを単一又は一体的なモデルで学習・推論するアーキテクチャをいう。

なお、End-to-End という用語が指す範囲は対象分野や利用文脈によって異なり、マルチモーダル AI 等では、画像・音声・テキスト等の認識処理における入力から出力までを指し、ロボットの判断・制御までを含まない場合がある。

VLM（Vision Language Model）

視覚言語モデル。画像や動画などの視覚に係る情報とテキスト等の言語情報を同時に処理できる AI モデル。

VLA（Vision Language Action）

視覚言語行動モデル。視覚、言語に加えて、動作や行動の 3 つのモダリティからの情報を統合的に処理する AI モデル。

フィジカル AI

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。サイバー空間での処理にとどまらず、アクチュエータ等を介して現実世界と直接相互作用し、移動、操作、加工、案内、搬送等の物理的な働きかけを行うことを特徴とする。

機能安全

EUC（Equipment Under Control：制御対象となる装置又は設備）及び EUC 制御系の全体に関する安全のうち、電気・電子・プログラマブル電子（E/E/PE）安全関連系及びその他のリスク低減手段の正常な機能に依存する部分。

なお、機能安全においては、対象設備・制御対象に係る危険源及びリスクを踏まえ、必要な安全機能及び安全度を定める[5]。

機械安全

[4] AISI「AI セーフティに関する評価観点ガイド（第 1.10 版）」（2025 年 3 月）
https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.10_ja.pdf

[5] IEC 61508-4 Functional safety of electrical/electronic/programmable electronic safety-related systems - Part 4: Definitions and abbreviations

機械の使用に伴う危険源を特定し、リスクアセスメント及びリスク低減を通じて、設計上の方策、安全防護、付加保護方策、使用上の情報等により、リスクを許容可能な水準まで低減する考え方又は取組み[6]。

[6] ISO 12100 Safety of machinery — General principles for design — Risk assessment and risk reduction

2. AI 利活用・セーフティの動向

2.1. AI ロボティクスの技術動向

従来、ロボティクスは製造業における組立やピッキング等に活用されることが主流であったが、AI の進展とその搭載により、AI ロボティクス領域として急速に発展している。生成 AI や VLM などのマルチモーダル生成 AI の導入によって、ロボットが認識・判断・行動を支援するなど、これまでにない利用者体験が実現されつつある。近年では、ロボットに特化したロボット基盤モデルの開発も進展している。大規模かつ多様なデータを事前に学習するロボット基盤モデルにより、様々なタスクに必要となるスキルを一般化することが可能となり、ロボットが未知環境において動作することが期待されている。ロボット基盤モデルの中でも、特に自然言語指示や視覚認識を物理的な行動に変換する VLA モデルの研究が盛んに取り組まれている。VLA を軸に、模倣学習や強化学習、実環境データとシミュレーションデータの併用によって、認識・判断・行動をより End to End で扱う発想が広がっている。さらに、実世界環境から得られる観測情報からその構造を学習によって獲得する世界モデル[7]への期待も高まっており、世界的に研究が進められている。こうしたソフトウェア面における AI の進展と、センサ等ハードウェア面の高度化が融合し、AI ロボティクスは急速な発展を見せている。ロボティクスは従来、産業用ロボットとしての活用が主流であったが、AI の活用を含む高度化に伴い適用領域が拡大し、人との協調の下で様々なタスクを実行することができるサービスロボットが、今後主流になると見込まれている。さらにその先には、自然言語による指示や視覚認識に基づいて行動に移す能力を持つフィジカル AI を搭載し、現実世界の物理法則を理解したうえで汎用的かつ自律的に行動するヒューマノイドロボットの社会実装が期待されている。

2.2. AI ロボティクスの市場動向

国内では 1990 年代から 2020 年頃にかけ、産業用ロボットのロボット密度（従業員 1 万人当たりのロボット数）がほぼ横ばいであった。これは市場が成熟期であったことを意味するが、その後、人間と協力して動作可能なロボットが実用に至ったことで、2020 年以降、新規のロボット導入が進みつつある。国内における AI ロボティクスの市場規模は 2030 年から 2050 年にかけて 2.5 兆円規模から 6.0 兆円規模まで成長すると推計されている[8]。2.1 に示したとおり、AI ロボティクスの活用は、サービスロボットの比重が高まる可能性があり、様々な産業分野への導入が見込ま

[7] 松尾・岩澤研究室「世界モデル」とは何か？知能の実現に向けて、松尾研が研究を推進する理由。https://weblab.t.u-tokyo.ac.jp/news/20221130/ - 閲覧日: 2026 年 6 月 2 日

[8] 三菱総合研究所 研究提言レポート「未来を切り拓く AI ロボティクス 日本がリードするための 3 つの提言」（2025 年 10 月）
https://www.mri.co.jp/knowledge/column/i5inlu000002n800-att/nr20251006.pdf

れている。例えば、物流やホスピタリティ、土木・建設、医療・介護、清掃、農業といった分野で、国内外を問わずAIロボティクスの導入が進みつつある。具体的には、ピッキング・搬送、配膳、清掃、警備・点検、建設現場での穴あけ・資材搬送、農作物の収穫といった特定作業に活用されている。ロボットの形態も、アームロボット、自律走行ロボット、四足歩行ロボット、モバイルマニピュレータ、ヒューマノイドロボットなど多様化しており、従来の製造現場に閉じた産業用ロボットから、人と同じ空間で移動・作業するサービスロボットへと適用領域が広がっている。一方で、高齢化や労働力不足を背景に、製造分野における産業用ロボットへの需要も大きく残ることが予想されている。

世界におけるAIロボティクスの市場規模は、2030年から2050年にかけ、27兆円規模から90兆円規模になると予想されている。AI ロボティクス分野の開発・実装においては、米国、中国、欧州が世界をリードしており、日本が強みを有してきた産業用ロボット市場の成長が横ばいの予測である一方で、サービスロボット市場は今後も拡大が予測されている。今後AIの搭載による自律化、汎用化が進むことで、国内外のサービスロボット市場の開発競争は過熱する状況になると予想される[9]ものの、主要各国の方向性は異なる。

米国ではビッグテック企業を中心に、民間主導での市場創出が進み、巨額の民間リスクマネーの動きがある。政府としての補助事業も実施されており、米国国立科学財団（National Science Foundation: NSF）によるAIロボティクスへの投資や、エネルギー省によるAIへの投資などが行われている。エネルギー省による AI 投資では、4 つのイニシアティブのうちの 1 つとして「Robotics and automation」が掲げられている[10]。

欧州では、EU AI法の施行など法的枠組みの確立を強化しており、規制を通じた競争優位を確立する方向性がある。その他、2021年6月には、「AI, Data and Robotics Association (Adra, asbl)」と欧州委員会がパートナーシップを締結し、EUの研究・イノベーション促進プログラム「Horizon Europe」による公的投資と、補完的な民間投資を通じて、2021年から2030年までにAI・データ・ロボティクス分野へ累計26億ユーロを投入するとしている。

中国では、第14次5カ年計画において「ロボット産業発展計画」を定め、国家主導の下で研究開発への巨額の投資が進んでいる。また、AIやロボティクス関連の人材育成に力を入れている点も特徴である。

急速な拡大が見込まれるAIロボティクス市場であるが、フィジカルAIの進展・普及による巨大な新規市場の形成も見込まれている。ヒューマノイドを含む多用途ロボットの市場規模は、2030年を境に急速に拡大し、2040年時点で60兆円規模になるとの見込みがある[11]。特に、巨額の投資が進んでいる米国や中国では、ヒューマノイドへの投資も顕著であり、一部では商用化も始まっ

[9] 経済産業省「AIロボティクス検討会 参考資料」（2025年10月8日 とりまとめ）
https://www.meti.go.jp/shingikai/mono_info_service/ai_robotics/pdf/20251008_2.pdf

[10] U.S. Department of Energy “Energy Department Advances Investments in AI for Science”
https://www.energy.gov/articles/energy-department-advances-investments-ai-science - 閲覧日: 2026年6月2日

[11] McKinsey & Company “Will embodied AI Create robotic coworkers?”
https://www.mckinsey.com/industries/industrials/our-insights/will-embodied-ai-create-robotic-coworkers - 閲覧日: 2026年6月2日

ている。

2.3. 利活用事例

サービスロボットの普及を背景に、表 2-1 のとおり、様々な産業での AI ロボティクスの導入が進展している。現状では、ピッキングや運搬、検査といった特定のタスクを高効率・高信頼に実行する特化型ロボットの活用事例がほとんどであるが、一部、多用途に対応可能なヒューマノイドの実用化が始まっている。

表 2-1　AI ロボティクスの利活用事例

産業カテゴリ	利活用事例
製造	自律移動、ピッキングなどを行うアームロボット
	非構造化環境で対象を認識しピッキングを行うアームロボット
物流・倉庫・輸送	車両からコンベアへの荷下ろしを行うアームロボット
	庫内輸送や作業者との協調作業を行う倉庫自動化ロボット
	様々な施設や場面で荷物を運ぶ自律走行ロボット
	飲食店で配膳や下げ膳、来客の案内を行う自律移動ロボット
	スーパー等で自律移動し、在庫状況や価格データ等を取得するロボット
	ホテル等で食事やリネン、清掃用品の配達を行う配送ロボット
医療・介護	病院での薬や用品の運搬、物品の取得といった患者対応以外の業務をサポートするロボット
清掃	医療施設や宿泊施設、工場、空港、倉庫など様々な場面で清掃を行う自律走行ロボット
警備・点検	施設内を巡回し、点検や異常検知を行う自律走行ロボット
	屋内外の様々な地形に適応し、巡回警備を行う自律走行ロボット
	化学プラントといった危険環境において自律移動し、設備の検査を行う四足歩行ロボット
建設	持ち上げやビス打ち、溶接など多能工作業を行うロボット 資材を自律搬送するロボット
	天井などアクセスしにくい場所に複数の穴をあけるといった肉体的負担の大きい作業を代替するロボット
農業	作物と雑草を識別し、雑草のみを除草する自律走行ロボット
	ハウス内を巡回し、作物の収穫を行うロボット 収量を予測するロボット

2.4. AI セーフティの政策・規制・標準化動向

AI 技術の急速な進展に伴い、ロボティクス分野をはじめ様々な分野で導入が進む一方、AI による誤判断、予期せぬ挙動、不適切な情報生成や発話等、AI 特有の心理的・社会的リスクを含めて評価する必要が生じている。これらのリスクは、ビジネスや実業務への影響だけでなく、活用分野によっては人命に影響を与える可能性もあるため、リスクの適切なマネジメントが不可欠となる。

このように、ロボティクス分野においても「AI セーフティ」が重要となっており、各国では AI セーフティを確保しながら AI の利活用を促進する取り組みが既に進められている。

2.4.1. 各国の動向

日本国内においては、「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）が設立された。また、同年の 4 月には、経済産業省と総務省から発行された既存の AI 関連ガイドラインを統合・アップデートした「AI 事業者ガイドライン」が策定された。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）[12]が公布・一部施行され、同年 9 月に全面施行された。さらに 2025 年 12 月には、AI 推進法に基づく計画として、「人工知能基本計画」[13]が閣議決定された。人工知能基本計画では、信頼できる AI の実現に向けたデータの整備、基盤モデル及び評価基盤の開発の推進について示されている。

一方で、国際的な動向として欧州では、生成 AI を含む包括的な AI 規制として EU AI 法が、2024 年 5 月に EU 理事会で最終承認され、同年 8 月に発効した。当該法律は、リスクベースで AI を分類し、リスク分類ごとに要求事項と義務を定めている。同年 5 月には、社会的・経済的な利益とイノベーションを促進しつつ、リスクを軽減する形で AI の開発、展開、利用を可能にすることを目指す「AI Office」を欧州委員会内に設立することが発表された[14]。さらに、2025 年 7 月には、汎用 AI モデルの①透明性、②著作権、③安全性・セキュリティに関する EU AI 法の遵守を支援する行動規範「The General-Purpose AI Code of Practice」が発表された。

同様に英国においても、AI（規制）法案（Artificial Intelligence (Regulation) Bill）が 2025 年 3 月に上院に提出され、AI 開発を包括的に監督する規制機関の設立が提案されている。また、英国は世界初の AI 安全専門機関として、UK AI セーフティ・インスティテュート（現 AI セキュリティ・インスティテュート：UK AISI）を 2023 年 11 月に設立した。UK AISI は、AI モデルの出力や挙動を解析できるツールとして、AI セーフティの評価プラットフォーム「Inspect」を公開して

---

[12] 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（令和七年法律第五十三号）
https://laws.e-gov.go.jp/law/507AC0000000053

[13] 人工知能基本計画 ～「信頼できるＡＩ」による「日本再起」～, 2025 年 12 月 23 日 閣議決定
https://www8.cao.go.jp/cstp/ai/ai_plan/aiplan_20251223.pdf

[14] European Commission, Commission establishes AI Office to strengthen EU leadership in safe and trustworthy Artificial Intelligence
https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2982 - 閲覧日: 2026 年 6 月 2 日

おり、AI コミュニティが自由に利用・改良できるようオープンソースとして提供している。一方で、UK AISI においては、2025 年 2 月に組織名を「AI Safety Institute」から「AI Security Institute」に改名することを発表し、改名に伴い、「バイアスや言論の自由等は扱わず、AI がもたらし得る最も深刻なリスクに注力」し、国家安全保障や犯罪悪用リスクへの対策を中心に取り組む方針が示された。2026 年 2 月には、UK AISI のホームページにおいて、先端 AI が国境・産業・言語をまたいで開発・利用される中で、信頼ある導入を支えるために共通の測定・評価の方法が必要であると指摘している[15]。

一方で、米国はこれらの制度とは対照的に、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いる。また、従来の U.S. AI セーフティ・インスティテュート（U.S. AI Safety Institute）を「AI 標準・イノベーションセンター（Center for AI Standards and Innovation: CAISI）」に改編した[16]。さらに、2025 年 7 月に米国の AI 競争力強化を狙いとした国家戦略「America's AI Action Plan」を発表した。当該国家戦略は、①AI イノベーションの促進、②AI インフラの構築、③国際的 AI 覇権と安全保障の確保を 3 本柱とし、特に①では過度な AI 規制を課す州への補助金制限やオープンソース型の AI モデルの推進等の政策方針を示しており、規制緩和の動きが高まっている。2025 年 12 月には、NIST のホームページにおいて、「AI の性能やリスクを適切に評価するためには、何をどのように測るかという測定科学が不可欠であるが、現在の AI 評価は未成熟であり、信頼性の高い評価手法の確立が必要となるため、AI 評価はサイバーセキュリティなどの具体的分野において実施され、それぞれに応じた厳密な測定手法の研究が不可欠である」と指摘されている[17]。

中国では、国家戦略として AI 産業の発展を促進しつつ、既存の法令と AI 関連法規制及び標準などを組み合わせる形で AI 規制を行っており、早期の全面的な AI 促進政策から、AI の倫理・安全、アルゴリズム規制などの法律を制定するようになっている。2017 年に国務院（中央政府）が公表した「新一代人工知能発展規画」[18]では、2025 年までに人工知能に関する基本的な法律、倫理枠組み、人工知能に対する安全評価と管理能力を構築するとしている。主な法律・規制としては、「インターネット情報サービスアルゴリズム推薦管理規定」（2021 年）、「生成 AI サービス利用暫定弁法」（2023 年）、「インターネット情報サービスにおけるディープ・シンセシス管理規定」（2023 年）、「人工智能科技倫理審査とサービス弁法（試行）」（2026 年）などがある[19]。また、2023

---

[15] UK AISI, International consensus and open questions in AI evaluations
https://www.aisi.gov.uk/blog/international-ai-network-consensus-and-open-questions - 閲覧日: 2026 年 6 月 2 日

[16] CRDS, 米商務省、AI 安全研究所を AI 標準・イノベーションセンターに再編
https://crds.jst.go.jp/dw/20250627/2025062742192/ - 閲覧日: 2026 年 6 月 2 日

[17] NIST, Accelerating AI Innovation Through Measurement Science
https://www.nist.gov/blogs/caisi-research-blog/accelerating-ai-innovation-through-measurement-science - 閲覧日: 2026 年 6 月 2 日

[18] 国务院关于印发新一代人工智能发展规划的通知_科技_中国政府网, 2017 年 7 月 8 日,
https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm - 閲覧日: 2026 年 6 月 2 日

[19] 内閣府 ＡＩ制度研究会（第２回），桃尾・松尾・難波法律事務所 松尾剛行弁護士　発表資料
https://www8.cao.go.jp/cstp/ai/ai_kenkyu/2kai/shiryou5.pdf

年から 3 年連続で AI 法の立法の推進が年度立法計画に組み込まれており、今後も推進と規制の両面での国家的な動きが進むと考えられる。

表 2-2　主要国の AI 制度や AI セーフティ関連の動向

国	政策・法律等	指針・ガイドライン	関連する取り組み・動向
日本	AI 推進法　（2025 年）	AI 事業者ガイドライン（2024 年、最新改定 2026 年） 人工知能基本計画（2025 年）	AISI における事業実証 WG の新設。
欧州	EU AI 法（2024 年）	汎用 AI の行動規範（General-Purpose AI Code of Practice）　（2025 年）	欧州委員会内に「AI Office」を設置
英国	AI（規制）法案（2025 年上院提出）	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）	AI Safety Institute が AI Security Institute に組織名を改名し、方針転換。 AI モデル評価プラットフォーム「Inspect」の公開。
米国	America’s AI Action Plan（2025 年）	AI リスクマネジメントフレームワーク（AI RMF）　（2023 年）	U.S. AI Safety Institute が CAISI に改編。
中国	新一代人工知能発展規画（2017 年） 人工智能科技倫理審査とサービス弁法（試行）（2026 年） インターネット情報サービスアルゴリズム推薦管理規定（2021 年） 生成 AI サービス利用暫定弁法（2023 年） インターネット情報サービスにおけるディープ・シンセシス管理規定（2023 年）	国家人工知能産業総合標準化システム構築ガイドライン（2024 年） 人工知能セキュリティガバナンスフレームワーク（2024 年）	AI 法の立法の推進を年度立法計画に 3 年連続で組み入れ

2.4.2. AI の信頼性とリスクに関する国際標準化の動向

ISO 及び IEC の合同専門委員会（JTC 1）の AI に関する分科会 SC 42 の WG3- Trustworthiness では、AI を社会的に「信頼できる」形で導入するための標準化が体系的に議論されている。また ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of

trustworthiness in artificial intelligence では、以下を包括的に定義している。

- 信頼性（Reliability）
- 可用性（Availability）
- 完全性（Integrity）
- 安全性（Safety）
- 説明可能性（Explainability）
- 透明性（Transparency）
- 説明責任（Accountability）

さらに WG3 は、SC42 で議論されている標準がこれらの信頼性のどの特性に関わるかを示すために、Trustworthiness Characteristics Matrix（TCM）を作成している。TCM は AI 関連標準開発において整合性確保とギャップ分析に活用され、ロードマップとともに随時更新されている。

WG3 が開発した、または開発中の主な成果を以下に示す。全体像、堅牢性、ソフトウェア品質(SQuaRE)、社会性と倫理、信頼性評価、HMT、人による監視などが挙げられる。

- ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
- ISO/IEC TR 24029-1:2021 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
- ISO/IEC 25059:2023 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
- ISO/IEC CD TS 22443 Information technology — Artificial intelligence — Guidance on addressing societal concerns and ethical considerations
- ISO/IEC AWI TS 25570 Information Technology — Artificial Intelligence — Reliability assessment of AI systems
- ISO/IEC CD TR 42109 Information technology — Artificial intelligence — Use cases of human-machine teaming
- ISO/IEC DIS 42105 Information technology — Artificial intelligence — Guidance for human oversight of AI systems

2.5. AI ロボティクスの政策・規制・標準化動向

AI セーフティに加えて、AI ロボティクスに関しても政策・規制に関する検討が各国で進んでいる。

2.5.1. 各国の動向

(1) 日本

AI のイノベーションを促進しつつリスクに対応するために、2025 年 6 月に公布・一部施行された AI 推進法、AI 推進法に基づく計画として 2025 年 12 月に閣議決定された、「人工知能基本計画」において、「信頼できる AI」の利用、イノベーション促進とリスク対応の両立等の基本方針が示されている。特に、フィジカル AI や AI ロボットは、AI 開発力の戦略的強化の一分野として位置づけられており、AI ロボットの公的需要創出や、より高度な自動運転技術導入を含めたフィジカル AI の研究開発及び実証を戦略的かつ統合的に促進することが具体的な取り組みとして示されている。

1.3.3 で紹介した「AI ロボティクス戦略」では、プライバシー、セーフティ、セキュリティの確保や、人とロボットとの協働を両立する観点から、必要な技術要件・基準の検証・整備を進めるとともに、高度な検証を行う体制に裏打ちされた安全性認証制度や安全規制の在り方を検討するとしている。

産業界では、日本ロボット工業会（JARA）が 2025 年 5 月に公表した「ロボット産業ビジョン 2050」[20]において、ロボット産業の将来像と課題が整理されている。同ビジョンにおいても AI に関する記述が多くみられ、ロボットと人間の共生における物理的な影響だけではなく、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についても考慮すべき問題として示されている。こうした状況を踏まえ、AI ロボティクスを含む先端技術の健全な普及のためには、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を策定することが重要であるとされている。

こうした AI 技術の進化に伴い、一般社団法人 AI ロボット協会（AIRoA）が 2024 年 12 月に設立され、2025 年より活動を本格化している[21]。AI とロボット技術の融合によるロボットデータエコシステムの構築を目指すとされており、産業の垣根を超えたオープンかつ大規模なデータ収集と統合、基盤モデル開発・オープンソース化、スケール可能な AI ロボットのエコシステムの構築、日本発のスタートアップや研究機関の支援・連携促進を主な取り組みとして推進するとしている。事業としては、AI ロボットの社会普及のための取り組みの一環として AI ロボットの安全性評価の検討・公開を行うとしている。

日本国内における AI ロボットの近年の取り組みの特徴として、ロボットによる物理的影響のみならず、AI ロボットを心理的・社会的影響を含むシステムとして捉え、責任分担とガバナンス考慮した社会実装を進める動きが進んでいる。

(2)　米国

[20] 一般社団法人日本ロボット工業会「ロボット産業ビジョン 2050 ― 人・社会・環境と共存するロボット ― Ver1.0」（2025 年 5 月）
https://www.jara.jp/publications/img/vision/visionver1_booklet.pdf

[21] PR TIMES, 一般社団法人 AI ロボット協会（AIRoA）AI×ロボット分野で、ロボットデータエコシステム構築を目指し活動を開始（2025 年 3 月 7 日）
https://prtimes.jp/main/html/rd/p/000000001.000158322.html - 閲覧日: 2026 年 6 月 2 日

米国では、ロボティクス分野で研究実績を有する複数の大学[22]が中心となってロードマップを取りまとめている。2024 年 4 月に公開されたロードマップの最新版「A Roadmap for US Robotics: Robotics for a Better Tomorrow」[23]では、国際競争力の確保のため、ロボティクス分野を国家・政府レベルでの最優先課題としたうえで、産業界・学術界・政府横断で取り組み、包括的なビジョンを定めるべきだと提言している。ロードマップにおいては、AI ロボティクスの実装において考慮すべき社会的観点として、倫理的な活用、包摂性の促進、安全性及びセキュリティの担保などが挙げられている。具体的な内容を以下に示す。

表 2-3　AI ロボティクスの実装において考慮すべき社会的観点

社会的観点	詳細項目
倫理的な活用 （Ensuring ethical application）	・意思決定の健全性を確保するためのガイドラインや標準の策定 ・データセットにおけるプライバシーの保護 ・AI の挙動における明確な説明責任の所在の確立
包摂性の促進 （Fostering inclusion）	・身体的・認知的能力の多様性を考慮した設計 ・情報格差（デジタルディバイド）の拡大防止 ・幅広い視点やバックグラウンドを持つ人々による AI 開発による、無意識の思い込み（アンコンシャスバイアス）を防ぐための取り組み
安全性の担保 （Guaranteeing safety）	・業務外で利用する一般ユーザ向けロボットの動作における身体的安全性の確保のための標準が不足していることへの対応 ・AI によってロボットを武器化することを禁止するための公共的な枠組みの策定 ・AI ロボットによる人間(特に子供の発達やメンタルヘルス、社会規範）への影響を考慮すること ・AI ロボットが高い信頼性及びフェイルセーフ機構を備えること
セキュリティの担保 （Guaranteeing security）	・安全保障の観点における標準の策定 ・政策によるサイバーセキュリティリスクへの対応 ・AI による過度な個人データ等の監視及び記録への規制

[22] University of California, San Diego, University of Pennsylvania, University of Texas, Austin, Arizona State University, Oregon State University, Stanford University, University of California, Irvine, University of Massachusetts, Lowell, University of Minnesota, Twin Cities, University of Southern California

[23] Congressional Robotics Caucus Advisory Committee,　A Roadmap for US Robotics: Robotics for a Better Tomorrow 2024 Edition（2025 年 3 月）
https://roboticscaucus.org/a-roadmap-for-us-robotics-robotics-for-a-better-tomorrow-2024-edition/　- 閲覧日: 2026 年 6 月 2 日

	・自律型兵器への AI の適用制限の必要性

産業界では、2025 年 7 月に米国の自動化推進団体であるオートメーション推進協会（Association for Advancing Automation：A3）が、産業界における AI の理解、評価、実装に関する詳細なロードマップを示すホワイトペーパー「AI in Automation: The Intelligent Transformation of Industry Today and Beyond」[24]を発表しており、ロボティクスを含む産業用途での AI 導入へ前向きな姿勢を示している。危険検知や異常予兆の把握など安全安心の確保そのものを AI 活用の重要な目的として位置づける動きも見られ、システム設計の初期段階から安全機能を設計に直接組み込む「Design for Automation（DfA）」の考え方が示されている。

(3)　欧州

欧州では、EU の研究・イノベーション促進プログラムとして、Horizon Europe 第 9 期（2021 年～2027 年）にて、「卓越した科学」、「グローバルな課題と欧州の産業競争力」、「イノベーティブ・ヨーロッパ」の 3 本柱を立て、総予算 955 億ユーロでの取り組みが行われている。その第 2 の柱「グローバルな課題と欧州の産業競争力」においては、（1）健康、（2）文化、創造性、包摂的な社会、（3）社会のための市民の安全、（4）デジタル、産業、宇宙、（5）気候、エネルギー、モビリティ、（6）食料、バイオエコノミー、資源、農業、環境、の 6 分野のクラスターが置かれている[25]。

2021 年 5 月にクラスター4「デジタル、産業、宇宙」における AI、データ、ロボティクスパートナーシップの民間部門として Adra が発足した。このパートナーシップは、AI、データ、ロボティクス分野における新たな市場や応用分野を拡大し、投資を呼び込むことで、企業、市民、環境に対して技術的・経済的・社会的価値を創出することを目指している。会員向けには Generative AI for robotics や Generative AI for manufacturing をはじめとするトピックグループが設置されており、リスクの低減も含めた AI の実装可能性を検討している。2024 年 12 月に Adra が発表した「Policy Paper and Technology Roadmap: GenAI and Robotics 4EU[26]」では、生成 AI はロボットに対して高度な学習能力、知覚能力、自然な対話能力を付与する一方で、安全性、説明可能性、サイバーセキュリティ、持続可能性といった重要課題に対処することが、欧州の価値観と長期的なイノベーションとの整合性を確保するうえで不可欠であると述べられている。

(4)　中国

---

[24] Association for advancing automation, Artificial Intelligence Whitepapers（2025 年 7 月）https://www.automate.org/ai/whitepapers/ai-in-automation-the-intelligent-transformation-of-industry-today-and-beyond - 閲覧日: 2026 年 6 月 2 日

[25] European Commission, Horizon Europe https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en - 閲覧日: 2026 年 6 月 2 日

[26] Adra “Policy Paper and Technology Roadmap: GenAI and Robotics 4EU” https://european-big-data-value-forum.eu/wp-content/uploads/2024/10/THE-Policy-Paper-and-Technology-Roadmap-GenAIRobotics-PetraK.D.pdf

中国では 2025 年 9 月に「人工智能安全治理框架 2.0（AI Safety Governance Framework 2.0）」[27]が発表された。2024 年 9 月に 1.0 版が発表された後、国家インターネット情報弁公室の指導の下、国家インターネット緊急対応センターを中心に、AI 専門機関、研究機関、業界企業をまとめ、2.0 版を策定した。2.0 版は 1.0 版をベースに、AI 技術の発展と応用の実践を結び付け、リスクの変化を継続的に追跡しながら、リスク分類を整備・最適化やリスクの階層化を検討し、防備・ガバナンス措置を動的に調整・更新している。

特に、AI ロボティクスに関連して、Embodied AI における技術の進展がデジタル知能と物理世界のラストワンマイルをつなぎ、人間と機械の統合が現実化すると同時に、AI セーフティリスクに対する顕在的影響及び認識が急速に進化していると指摘されている。具体的には、AI による現実世界及び認知的リスク、差別的バイアス、公共利益や個人への権利侵害などの社会的影響に加え、心理的影響として AI の予測不可能かつ追跡不能な意思決定が人間に与える影響についての言及がなされている。

また、2025 年 12 月には中国国家インターネット情報弁公室が、「AI を擬人化した対話サービスの管理に関する暫定規則」の草案（意見募集稿）を公表した[28]。AI を擬人化した対話サービスとは、「AI 技術を用いて、人間の性格特性、思考パターン、コミュニケーションスタイルをシミュレートし、テキスト、画像、音声、ビデオその他の手段を通じて人間と感情的に交流する製品やサービス」と定義されている。草案の中では、これらが応用されると考えられるロボット及びスマートシステム向けの具体的な 24 の規則として、AI システムが人間らしさを持ち、人間に対して心理的影響を与え得ることを想定し、心理的安全や依存リスク、未成年者や高齢者の保護、感情操作のリスクなどが挙げられている。またこれらの規則に従い行政に対する定期的な書面の提出や、監査の実施も規定されている。

2.5.2. AI システムと機能安全の国際標準化の動向

ISO/IEC JTC 1/SC 42 JWG 4 - AI and Functional Safety では、AI と機能安全の議論を行っており、ISO/IEC TS 22440-1 Artificial intelligence — Functional safety and AI systems — Part 1: Requirements の開発を進めている。この規格は ISO/IEC TR5469:2024 の後続規格として策定されている。

機能安全の考え方は、安全制御の確実性を担保するために、回路故障や手法不具合を排除することにある。AI 固有の不具合や不確実性を分析し、制約を加え、追加の確認を行うことで、AI 技術を安全機能に使うことについて検討を進めている。使用する AI ソフトウェア技術ごとに、実現する安全プログラムのリスクレベル(Application Usage Level: AUL)に応じて、制限や追加手法を

[27] 中央网络安全和信息化委员会办公室，《人工智能安全治理框架》2.0 版（2025 年 9 月）
https://www.cac.gov.cn/2025-09/15/c_1759653448369123.htm - 閲覧日: 2026 年 6 月 2 日

[28] 国家互联网信息办公室，国家互联网信息办公室关于《人工智能拟人化互动服务管理暂行办法（征求意见稿）》公开征求意见的通知（公表日 2025 年 12 月 27 日、意見提出期限 2026 年 1 月 25 日）
https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm - 閲覧日: 2026 年 6 月 2 日

規定している。特に、AI 固有の問題である学習データ及び学習における確実性確保について、議論が続いている。

また、AI を使わないロボットの機能安全の一例として、開発ツールに AI 技術を使用する場合の制限についても議論が行われており、現状の安全制御システムの安全プログラムを、生成 AI により作成する場合などが考えられる。JWG 4 で開発中の成果物には以下がある。Part2 では、安全関連機能における AI 技術の利用や AI 技術を用いたシステムの安全性を確保するための従来技術に基づく安全関連機能の利用等に関する指針、Part3 ではそれらに関する適用例が示される予定である。

- ISO/IEC AWI TS 22440-2 Artificial intelligence — Functional safety and AI systems — Part 2: Guidance
- ISO/IEC AWI TS 22440-3 Artificial intelligence — Functional safety and AI systems — Part 3: Examples of application

当該報告書では、①安全機能そのものに AI を利用する場合、②AI の誤動作を非 AI 技術で補完する場合、③安全関連システムの設計・開発に AI を利用する場合、の 3 類型を整理している。また、AI 特有のリスク要因、品質確認方法、既存の機能安全規格との整合性などを示し、安全性確保の指針を提供している。

3. AI ロボットのリスクと要因

本章では、AI ロボットについて、社会実装を進めるうえで想定されるリスクとその要因を整理する。AI ロボットは、センサ入力から行動（移動・把持・対話等）に至る意思決定が高度化することによりブラックボックス化しやすく、従来の機械安全や機能安全だけでは捉えきれないリスクが顕在化し得るため、物理的な安全系のリスクに加え新たな類型で整理する必要がある。本書で AI ロボットを対象として示したリスクは、フィジカル AI や End to End モデルの社会実装においても同様に想定されるリスク分類として整理を目指したものであり、本章の議論はそれらに関する本質的な論点を概ね包含したものとなっている。

3.1. 想定されるリスク

AI ロボットに関する想定リスクを、大きく以下の 3 類型に分類する。これら 3 類型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。)

① 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)。

② 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)。

③ 社会的リスク
サイバーセキュリティ攻撃や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 3-1　想定リスクの類型

3.1.1. 物理的リスク

AI ロボットの物理的リスクは、従来から産業用ロボット等で議論されてきた挟圧・巻き込み・衝突等に加え、AI による認識・判断・制御の不確実性が上乗せされる点に特徴がある。表 3-1 は、物理的リスクの例だが、特に、人と空間を共有する協働ロボット、物流ロボット、サービスロボット等では、環境変動（人の動線、物体配置、照度、騒音等）が大きく、想定外挙動の誘因となり得る。

表 3-1　物理的リスクの例

リスク例	事例・研究例
・人を物体と誤認識し事故が発生 ・経路計画の失敗による人とのニアミス、接触事故 ・ヒューマノイドロボットの転倒や誤動作による人への直撃	農作物流通センターでロボットが人間を箱と誤認して圧死させる事故が発生[29]
・物体を誤って把持し損壊 ・物品の把持ミスで落下させ人に直撃 ・家庭や店内での食器等、物品・商品破損	エアー駆動のスイングロボットが成型機上で作業を行う際、工具が滑って成型機上から落下し金型を損傷[30]
・AI が「学習外の状況」で想定外の動作を取る ・指示文の誤解釈で不適切な高さ・速度・力で動作する ・ロボット同士の協調制御の乱れによる誤動作	AI ヒューマノイドが発表会にて転倒[31]
・非接触センサ（LiDAR、カメラ）の故障による停止機能の未作動 ・インターロックや緊急停止の誤動作・無効化	ヒューマノイドロボットが突然腕を振り回して、作業員が負傷[32]
・ロボットの能力を過信し、危険領域に接近し動作中のロボットに衝突	自律配送ロボットが歩道で歩行者と接触（歩行者側は「ロボットは自分を検知して避けてくれるはず」と無意識に期待）[33]

[29] The Guardian, Industrial robot crushes man to death in South Korean distribution centre https://www.theguardian.com/technology/2023/nov/08/south-korean-man-killed-by-industrial-robot-in-distribution-centre - 閲覧日: 2026 年 6 月 2 日

[30] 株式会社ハーモ，射出成形の工程改善ガイド https://www.injection-molding.jp/ - 閲覧日: 2026 年 6 月 2 日

[31] Humanoid Robot Falls on Its Face While Making Debut as Company Says They're 'Puzzled' by Sense of Global Concern https://people.com/humanoid-robot-falls-on-face-company-puzzled-global-concern-11849102 - 閲覧日: 2026 年 6 月 2 日

[32] Humanoid robot malfunctions, hits worker in China– is AI getting too real? https://timesofindia.indiatimes.com/etimes/trending/watch-humanoid-robot-malfunctions-hits-worker-in-china-is-ai-getting-too-real/articleshow/120932185.cms

[33] Avride says its robot was safe and compliant when a pedestrian caused a collision in Austin https://jp.helm.news/2025-10-02/avride-says-its-robot-was-safe-compliant-when-pedestrian-caused-collision.html - 閲覧日: 2026 年 6 月 2 日

・高齢者・子供の予測困難な行動をロボットが適切に認識できない	

3.1.2. 心理的リスク

AI ロボットは、身体性（外観、動作、距離感）と対話（言語・声質・表情等）を通じて人間の心理面に影響を与える。

表 3-2 は、心理的リスクの例だが、物理的リスクと異なり「発生の検知」「被害の定量化」「責任の所在」が曖昧になりやすい。また、高齢者、子供、患者等の対象者、影響を受ける頻度、利用目的、利用状況等によってその影響の大きさは異なるため、利用目的、利用環境、対象利用者属性等を踏まえ、合理的に予見可能な影響を中心に整理した。

表 3-2　心理的リスクの例

リスク例	事例・研究例
・生成 AI が不適切・攻撃的・高圧的な言語を生成 ・しつこい勧誘・提案など過度な対話インタラクション ・誤った叱責・注意など、利用者が不快になる応答	AI セラピーチャットボットの不適切応答[34]
・顔認識の誤り（別人扱い、年齢・性別誤認） ・言ったことを理解してもらえないいらだち ・不安・混乱を感じさせる誤応答	高齢者ケアにおけるコミュニケーションロボット運用でのミスコミュニケーション[35]
・人とロボットの「共同作業」が増えることにより、ロボットへの過度な依存からくるヒューマンエラーが増加 ・ロボットを「人と同等」に扱い、過度に依存 ・ロボットに過剰に依存し、家族・友人との交流が減少したことによる関係悪化 ・子供が「友達の代替」として過度に依存	長期間使っていたコミュニケーションロボット喪失後の心理影響[36] 高齢者施設で孤独感低減やコミュニケーション促進に貢献する一方、過度に依存すると、人との交流時間が削られる懸念[37]

[34] Exploring the Dangers of AI in Mental Health Care, https://hai.stanford.edu/news/exploring-the-dangers-of-ai-in-mental-health-care - 閲覧日: 2026 年 6 月 2 日

[35] 太田 一実, 滝沢 龍「高齢者・認知症ケアへのコミュニケーションロボットの活用に関する文献レビュー」https://www.jstage.jst.go.jp/article/sjpr/65/4/65_395/_article/-char/ja/

[36] Long-term effect of the absence of a companion robot on older adults: A preliminary pilot study https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2023.1129506/full - 閲覧日: 2026 年 6 月 2 日

[37] The impact of care robots on older adults: A systematic review https://www.sciencedirect.com/science/article/abs/pii/S0197457225003507 - 閲覧日: 2026 年 6 月 2 日

・「感情を持っているように見える」振る舞いによる心理誘導 ・有名人等の個人データ（表情・声色）を用いたパーソナライズで、購買行動を誘導 ・高齢者・子供など脆弱層に対する影響	ロボットが、同情や愛着を引き出すような振る舞いを行うことで、望ましくない意思決定（不必要な購買、データ提供等）を誘導する危険性[38]

3.1.3. 社会的リスク

AI ロボットは、個人のみならず、企業を含む組織やコミュニティに対しても、利用形態や適用分野によっては、個別事故を超える社会的影響が生じる可能性がある。

表 3-3 の社会的リスクは、個人に係る差別・偏見、プライバシーの問題に加え、雇用や企業活動に係るセキュリティ、データ改ざんといった分類で整理する。

表 3-3　社会的リスクの例

リスク例	事例・研究例
・言語モデルが訓練データの偏見を反映し、差別的な発話 ・顔認識 AI が特定の属性（性別・年齢・民族）を誤識別 ・店舗・公共空間での顧客属性の誤判定による差別的扱い	LLM 搭載ロボットのシミュレーション実験で差別・暴力的行動[39]
・店舗・家庭でロボットが常時カメラ・マイクを稼働 ・会話・映像・行動データ、それらのログデータがクラウド上に意図せず蓄積	ロボット掃除機・芝刈りロボットがハッキングされ、カメラ・マイク等を遠隔操作される可能性[40]
・ロボットがサイバー攻撃で乗っ取られ、不適切な動作を行う ・脆弱性のあるソフトウェアがロボットに実装され、誤動作を起こす	配膳ロボットの認証トークンの問題などから外部から制御を奪える脆弱性[41]

[38] Mario Kropf “Mental Health, AI-based Care Robots and Fair Access to Healthcare” https://www.erudit.org/en/journals/bioethics/2025-v8-n3-bioethics010132/1118902ar.pdf - 閲覧日: 2026 年 6 月 2 日

[39] AI-Powered Robots Can Be Tricked Into Acts of Violence https://www.wired.com/story/researchers-llm-ai-robot-violence/ - 閲覧日: 2026 年 6 月 2 日

[40] Ecovacs home robots can be hacked to spy on their owners, researchers say, https://techcrunch.com/2024/08/09/ecovacs-home-robots-can-be-hacked-to-spy-on-their-owners-researchers-say/ - 閲覧日: 2026 年 6 月 2 日

[41] Researcher who found McDonald's free-food hack turns her attention to Chinese restaurant robots https://www.theregister.com/security/2025/08/29/chinese-pudu-robots-found-open-to-hijacking/946762?_gl=1*17igg4b*_ga*MTY0MzUyMTUwNS4xNzgwNDA4OTEz*_ga_JXW44Y23NM*czE3ODA0MDg5MTIkbzEkZzEkdDE3ODA0MDg5MjckajQ1JGwwJGgw -閲覧日: 2026 年 6 月 2 日

・サイバー攻撃で乗っ取られた店舗ロボットが偽情報を顧客に提供 ・正規オペレータになりすまして遠隔操作する	通信を改ざんし、ロボットの状態を偽装したり誤動作させたりする
・ロボットによる代替が特定職種に偏り、雇用の偏在を生む ・高額機器の導入ができる企業とできない企業の格差 ・自動化が進むと、社会的スキル（接客、介護コミュニケーション）が希薄化	倉庫などで、ロボットの導入により離職に追い込まれる

これらに加え、事業判断・契約・補償設計の観点では、事故やトラブルによる経済的影響についても考慮する必要がある。例えば、身体や物体等に直接与える影響を一次損失、操業停止・逸失利益・事業継続への影響・ブランド毀損等間接的な影響を二次損失として区別することが考えられる。このように整理することで、責任範囲（契約上の責任制限・免責を含む）や保険設計、事故原因分析及び立証に必要なログその他記録の管理等を検討するうえで重要となる。

ただし、経済的影響の大小のみに基づいて安全対策の要否を判断することは適切ではない。経済的影響の見積りに誤りがあった場合に、安全上の重要なリスクの見落としにつながり得るためである。そのため、経済的影響はあくまで検討要素の 1 つとして扱い、リスク低減の原則（合理的に実行可能な低減）の考え方を踏まえて安全対策を検討するべきである。

3.2. リスク発生要因について

AI ロボットのリスクは、従来から産業用ロボット等で議論されてきた人への挟圧・巻き込み・衝突等の物理的リスクに加えて、3.1 節で取り上げた心理的リスク、社会的リスクへの対応も考慮する必要がある。ここではリスクの発生要因として、「AI」、「ロボット」に加えて、AI 及びロボットを扱う「人間」、ならびにロボットを利用するための環境やルール等を含む「社会」を考える。

これら 4 つのリスク要因について、物理的／心理的／社会的リスクの事例を整理すると、以下のとおりとなる。

- AI

  物理的：指示文を誤解し、ロボットを想定外の速度や軌道で動作させる。

  心理的：高圧的・攻撃的な応答をする。

  社会的：AI により自動化された特定の職種が急速に代替される。

- ロボット

  物理的：センサ故障により急旋回・急停止する。

  心理的：ロボットの声質や表情が威圧的に感じられる。

  社会的：ロボットのログ・認証情報が悪意のある者に盗まれ乗っ取られる。

- 人間

  物理的：作業者がロボットの稼働域に不用意に侵入する。

  心理的：孤独の代替としてロボットに依存しすぎる。

  社会的：AI の差別的な判断結果を認知せず採用活動を行う。

- 社会

  物理的：歩道・公道でのロボット走行ルールが未整備であることにより事故が発生する。

  心理的：行動誘導の規制が未整備であること等により悪質なマーケティングが加速する。

  社会的：セーフティネット不足で格差が拡大する。

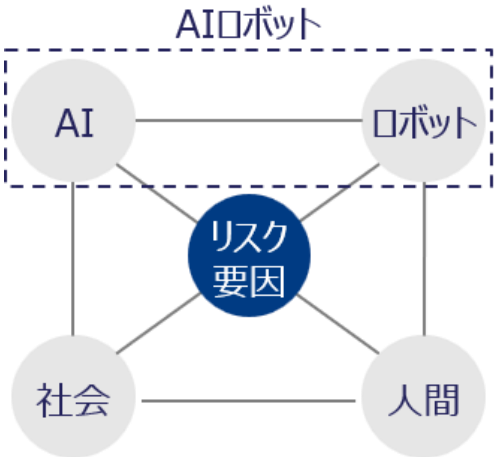


図 3-2　リスク要因

ロボットと AI それぞれの観点から考えると、従来の AI を搭載しないロボットにおいては、物理的リスクや社会的リスクの一部に対して、ISO 10218（産業用ロボットの安全規格）や ISO 13482（サービスロボットの安全要求事項）等で機能安全や機械安全について議論されている。他方、LLM 等の生成 AI では、テキストベースのアプリケーションを中心に、有害情報や偽誤情報、バイアス等による心理的リスクや社会的リスクへの影響が懸念されており、2.4 節で取り上げた各国の政策・規制をはじめ、ISO/IEC 42001（AI マネジメントシステム）等で議論されている。

つまり、ロボットが AI を搭載しそれらが組み合わさることで、ロボットを開発・提供する観点からは、物理的リスクに加えて、AI による心理的リスクや社会的リスクの懸念が、また AI を開発・提供する観点からは、心理的リスク・社会的リスクに加えて、物理的リスクの懸念が生じる。そのため、AI ロボットがもたらす便益に対して、AI を利用することによって新たに発生し得るリスクの要因を見極めることが重要になる。

3.3. AI ロボティクスのセーフティ評価の必要性

AI ロボティクス分野は、AI 利活用分野の中でも特に人の身体的安全と密接に関係する領域であり、従来から機械安全や機能安全の枠組みに基づく制度整備が進められてきた分野である。本節では、AI ロボティクスに関するリスクのうち、主として人の身体、物体、設備等に影響を及ぼし得る物理的リスクを中心に、従来の機械安全及び機能安全の観点から整理する。

産業用ロボットに代表される従来型ロボットにおいては、衝突や接触、感電・火災等の主として物理的危険源を対象とし、ハードウェア及び制御系の信頼性確保を中心とした安全設計・評価が行われてきた。具体的には、ISO 12100 等で示される機械安全の基本原則において、安全に関わるシステムの故障や誤りを極限まで減らすため、そうした安全機能の実現に必要なシステムのすべての部分、すなわち安全関連系を、それ以外の部分から独立させることが求められている。また、この安全関連系を設計・実装するにあたっては、IEC 61508 に安全関連システムが要求される SIL（Safety Integrity Level：安全度水準）を達成するために採用すべき開発プロセスや設計技法、検証手法、故障回避手法等が示されており、それらを参照することが望ましい。

これに対して、生成 AI や大規模言語モデル、マルチモーダル AI の進展により、ロボットが高度な自律判断機能や対話機能を備え、人間の近傍で行動するケースが増加している。こうした近年の状況においては、1.1 節に示したとおり、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となっている。こうした AI を、安全関連系を含むシステムに導入する場合、AI を安全関連系に導入するのか、あるいは非安全関連系に導入するのかにより、設計開発において考慮すべき点や実現方法に大きな違いが生じる。このうち安全関連系への AI の利用については、2.5.2 項に示した ISO/IEC JTC 1/SC 42 JWG 4 - AI and Functional Safety において議論が進んでいる。

「ロボット安全に基づいた人工知能安全三方針」では、図 3-3 のとおり、こうした二つの場合のポリシーの違いを、1st ポリシー、2nd ポリシーとして、従来の機械安全の原則と対比させて示

している。また、3rd ポリシーとして、機械システムに AI を導入するのではなく、従来の人の役割を代替するものとして AI が位置づけられる場合もある。

例えば、自動車において、ハンドルやブレーキ、アクセルなどの人の運転操作を正しく自動車の制御に結び付けるのが機能安全であるのに対して、人間の運転操作に関する指令自体を AI で代替する自動運転機能においては、AI を機能安全の中で扱うことが難しく、AI 自体の信頼性を高めたうえで、人間による運転操作の結果の比較等により評価せざるを得ない。このように、AI が人の役割を代替する用途では、1st ポリシー及び 2nd ポリシーと区別し、3rd ポリシーとして扱うことが妥当である。

このように、従来の機械安全の原則の観点から、AI のシステムへの導入については、これら 3 つのポリシーに基づいて検討することができるが、各ポリシーに応じて AI 及び機械システムの安全性の示し方を区別する必要がある。ただし本書では、これら各ポリシーにおける安全性の示し方詳細について整理することはせず、今後の更新時の課題とする。

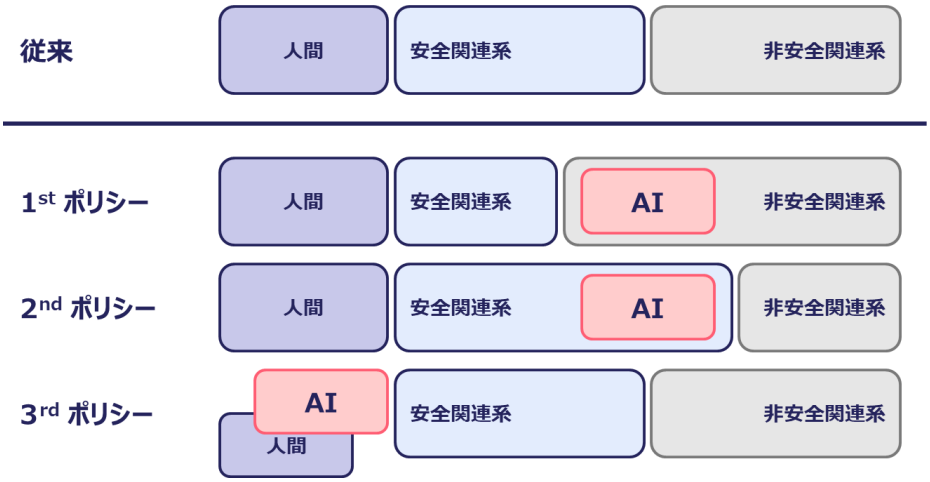


図 3-3　機械安全原則に基づく AI 導入の三方針[42]

なお、AI ロボティクスのセーフティ評価において考慮すべきリスクは、従来の機械安全や機能安全が主に対象としてきた物理的リスクに限らない。AI ロボットが人と対話し、人の近傍で行動し、又は人の判断・行動に影響を及ぼす場合には、3.1 節で述べたように心理的リスク、社会的リスクも考慮する必要がある。そのため、AI ロボティクスのセーフティ評価にあたっては、本節で整理した物理的リスクを中心とする観点に加え、心理的リスク及び社会的リスクも含めて、総合的に評価することが必要である。

[42] 藤原清司，角保志，尾暮拓也，中坊嘉宏，「ロボット安全に基づいた人工知能安全三方針」，日本機械学会ロボティクス・メカトロニクス講演会講演概要集（ROBOMECH2017），1A1-F01（2017 年 5 月）
https://www.jstage.jst.go.jp/article/jsmermd/2017/0/2017_1A1-F01/_pdf/-char/ja

4. AI ロボティクスに関するセーフティ評価観点

本章では、AI ロボティクスに関するセーフティ評価観点を示すとともに、各観点について、概要、想定されるリスク、評価項目例を整理する。AI ロボットの設計、開発、提供、運用、運用状況のチェックなどの各フェーズで参照することを想定する。

図 4-1 は、「AI ロボティクス」と「実世界環境」の関係を示している。AI ロボティクスは、周囲の環境や対象を捉える「認識」、行うべき動作や手順を決定する「計画・判断」、実際の動作を実行・調整する「制御・行動」が相互に関連している。その中で、リスク低減とタスク遂行を両立させることが求められる。実世界環境には、人、インフラ、社会環境などが含まれ、AI ロボットは、この環境の中で物理的な動作を行うだけでなく、人との関わりや社会制度との接点を持ちながら運用されるため、リスクは単一ではなく多層的に現れる。

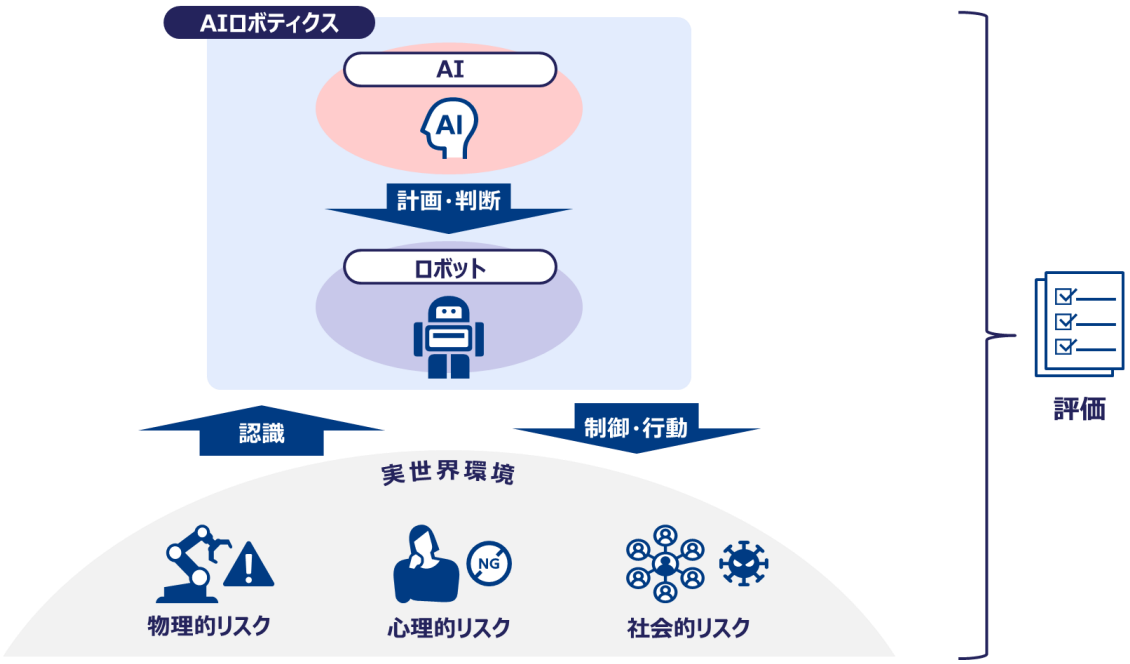


図 4-1　AI ロボティクスと実世界環境との関係イメージ

第 3 章では、AI ロボットの利用に伴って生じ得る物理的・心理的・社会的リスクと、その根拠となる考え方を整理した。法令、利用環境、想定利用者、社会的受容性等を踏まえ合理的に受容可能と判断される範囲までリスクの低減を図るために、開発や提供を行う事業者は対策を予防的に実施し、設計時評価に加え、運用時の監視、インシデント分析、継続的改善などを含めたリスク管理を実施する必要がある。

3.1 節で示した主要なリスク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に整理する。ただし、ここで整理した評価観点は、本書で想定する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洩の懸念、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洩の懸念を低減する**「プライバシーの保護」**
  - 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」**

以降の各節では、LLM を構成要素とする AI システム（LLM システム）を対象とした「AI セーフティに関する評価観点ガイド」の 3 章「評価観点の詳細」の表記スタイルに基づき、以下の項目を記載する。なお、「AI セーフティに関する評価観点ガイド」では、AI エージェントシステムの外部連携の一類型として、ロボット等を含む「物理世界操作」という整理がされている。本書は、この点も踏まえ、AI を搭載したロボットの開発・提供・運用において参照すべきセーフティ評価観点を示すものである。

以下の評価項目例は、評価時に確認すべき観点を例示するものである。個別ユースケースにおけるリスク低減の程度や判定基準は、対象ロボットの用途、運用環境、利用者、想定リスク等に応じて具体化することが想定される。

■ 評価観点の概要説明

評価観点の内容や、評価を通して目指す状態を説明する。

■ 想定され得るリスクの例

評価観点に関連して実際に生じ得るリスクの例や、将来的に生じ得るリスクの例を示す。あくまで例示であり、発生するリスクはここに列挙した例に限らないため、実際の環境に応じてリスクを検討することが重要である。

■ 評価項目例

各観点についての評価を行う場合に想定される評価項目の例を示す。あくまで例示であり、評価術項目はここに列挙した例に限らないため、実際の環境に応じてリスクを考慮し、必要な評価内容を検討することが重要である。

4.1. 対象ユースケースのセーフティ評価観点

4.1.1. 物理的安全性

■ 評価観点の概要説明

本観点は、物理的リスクに関連する。

認知・判断の結果に基づくロボットの移動や停止、把持・受け渡し等の動作が、物理的に安全であることを評価する。特に、移動時の速度や停止位置、把持の制御といった行動の適切性を確認する。

また、ロボットが意図した動作を行う正常時のみならず、センサ異常や通信遅延、予期せぬ障害物の出現といった想定外の事象における動作を含む。常に変化する環境を踏まえ、最適な解決策を適用しているかなど、行動の適切性を評価し、必要に応じて継続的に改善することが重要である。

■ 想定され得るリスクの例

- 過度な速度での移動や急停止により、物体や人間と接触し事故が発生する。
- 停止位置を誤り、出口や通路を塞ぎ、往来を妨げる。
- マニピュレータの把持や受け渡し動作における制御不備により、対象物の破損または人体に衝撃を与える。
- センサの死角（逆光・反射・遮蔽物を含む）で障害物を未検出のまま動作を継続することにより衝突する。
- 複数台のロボットを同時に運用する際、走行軌道が干渉し、ロボット同士または人との衝突が発生する。

■ 評価項目例

- 移動時の速度・停止タイミング
  - 人混み・狭路・交差点など状況別に速度上限が妥当に設定されている。
  - 人との距離に応じた制御（近接時の減速開始・停止）が実環境において機能する。

- 把持・受け渡し
  - 把持力の上限が対象物の材質・重量ごとに設定されている。
  - 物体の受け渡し時に相手の準備完了を確認してから動作を開始する手順が定義され、機能する。

4.1.2. 運用合理性

■ 評価観点の概要説明

本観点は、物理的リスク及び心理的リスクに関連する。

認知・判断の結果に基づくロボットの動作（移動・停止・回避・把持・受け渡し等）が、運用上において合理的であることを評価する。特に、経路切り替えや代替手段の提案といった行動の適切性を確認する。

具体的には、不必要な停止や過度な迂回、非効率な経路選択は、周囲の人が補助や介入を試みる誘因となり、二次的な事故リスクを生じさせる。周辺環境を問わずに走行できることを想定したものではなく、一定範囲に限定された個別の屋外・屋内環境に対して、動作が合理的であるかを確認するものである。

■ 想定され得るリスクの例

- 不必要に頻繁な停止や過度な迂回により事故を誘発する。
- 配達先が不在の状況で置き配に関する判断を誤り、荷物を誤った場所に放置する、または停止状態により通路を長時間塞ぐ。
- 狭路や交差点、非常口前など危険な位置で待機することで、通行者や緊急車両の妨げとなる。
- 軽微な障害物に対して大きく迂回することで予測不可能な軌道を生成し、周囲の人に接触する。

■ 評価項目例

- 適切な移動・停止
  - 商品等の受け渡し位置や待機位置、回避時の退避位置が安全かつ適切に選択される。
  - 停止位置が非常口・避難経路・設備操作パネル等の妨げにならないことを、導入環境ごとに確認する。
- 回避挙動
  - 対向・追い越し・横切りの各シナリオで安全距離が維持される。
  - 動的障害物（移動中の人・台車等）に対し、移動速度の予測に基づく先行的な回避が行われる。
- 代替提案の合理性
  - 配送ロボットのケースで、配送先が不在、施錠または経路が塞がっている等の状態に

おいて、置き配への切り替えやオペレータ呼出などの代替行動の選択肢が適切に定義されており、状況に応じて合理的な選択を行える。

- 代替提案の選択基準と実施結果がログに記録されており、事後検証が可能である。

- 例外対応・フェイルセーフ
  - 進路が塞がれた場合、一時停止を行い、場合によっては回避の協力を他者へ依頼する等の段階的な選択が妥当に機能する。
  - 緊急停止ボタンの応答時間、停止距離が規定値を満たしている。

4.1.3. 安定性

■ 評価観点の概要説明

本観点は、物理的リスク及び心理的リスクに関連する。

ロボットの移動、停止、姿勢保持等の物理的動作や振る舞いが安定しているかを評価するものである。本観点における安定性とは、転倒や把持したものの落下を起こさないという意味にとどまらず、移動時の加減速や経路・把持力などが適切に制御され、外乱や環境変動があっても、急激な挙動変化や予測困難な動き・発話を生じない状態を指す。

AI ロボティクスでは、認知・判断結果に基づき動作が動的に生成・更新されるため、認識の揺らぎや環境変化がそのまま挙動の不安定化につながる可能性がある。そのため、動作が滑らかで一貫性があり、周囲の人が予測可能で、安全側に収束する制御が確保された状態を目指す必要がある。

■ 想定され得るリスクの例

- 搬送ロボットの急激な加速により、積載物が落下する。
- 急加速・急停止・急旋回により、周辺の物体や人間と接触する。
- 床面段差や傾斜で転倒する。
- 把持力の不適切な制御により把持したものを破損したり、落下させたりする。
- 不規則な振る舞いや発話音量の不安定さにより、会話に支障をきたす。

■ 評価項目例

- 加減速・停止タイミング
  - ロボットの移動時において急な加減速が発生しない。
  - 効率的な経路で移動し、別経路に切り替えた場合でも、急激な加減速や旋回が生じない。
  - 非常停止時に姿勢崩れや横転を起こさない。
  - 停止距離が一貫している。
- 姿勢・重心
  - 最大積載時やアームが伸展する場合でも重心が安定している。

- 段差や傾斜がある床面でも、安定して自立できる。移動時も転倒しない。

- 把持・操作
  - 把持力が対象物の特性（割れ物等）に応じて適切に制御され、振動や落下が生じない。
- 振る舞い・動作
  - 例外発生時でも動作や発話音量等が安定している。

4.1.4. 人間の適切な介在

■ 評価観点の概要説明

本観点は、物理的リスク、心理的リスク及び社会的リスクに関連する。

必要な局面で人間（運用者・スタッフ・遠隔支援者）が介入できること、またロボット自身が不確実時に介入を適切に要請できることを評価する。介入は非常停止のような強制停止だけでなく、停止→協力依頼→遠隔支援といった段階的動作や、例外時の代替運用を含む。ロボットがAIにより高度化されても、人間の適切な介在ができているかを評価する。

■ 想定され得るリスクの例

- 人間の介入手段が分かりにくいまたは遅いことで、ロボットの不適切な挙動が事故につながる。
- 人間の介在が頻発する場合、異常時における介入判断や対応が形骸化するおそれがある。
- 介在のための遠隔操作の権限が攻撃者に奪われ、不正操作により危害が発生する。
- 人間の介在が終わった後の自律動作の再開条件が不適切または曖昧さがあることにより、介在すべき危険な状態が継続したままロボットが作動する。

■ 評価項目例

- 介在の適切な設計
  - 停止時・低速時・手動／遠隔操作切り替え時等、人間の介在タイミングが定義されている。
  - 進路塞がり等の場合に、「一時停止→協力依頼→必要に応じ遠隔支援」といった選択が適切に選べる。
- 権限管理
  - 介在することが可能な管理者が設定されている。
  - 介在時の認証や記録（誰がいつ何をしたか）が残り、ログとして一定期間保存される。
- 例外シナリオ
  - 荷物配送のユースケースで、配送先が不在または反応がない場合等、誤配送が疑われる場合に、オペレータ等へ連絡する、あるいは配送先住所の再確認等の代替手段を実行する。

4.1.5. プライバシー保護

■ 評価観点の概要説明

本観点は、社会的リスクに関連する。

ロボットが収集・生成・表示する情報（映像、音声、位置、行動履歴、カフェなどの注文内容等）については、収集する情報が必要最小限に限定されているか、利用目的の範囲内で適切に取り扱われているか、また必要な保護措置が実装されているかを確認する。あわせて、公共空間を含む実際の運用環境において、これらの保護措置が有効に機能しているかを評価する。

■ 想定され得るリスクの例

- 公共空間における音声読み上げや画面表示により、個人名や住所、電話番号等が周囲に露出してしまう。
- 映像・音声・行動データ等が意図せず蓄積されることにより、目的外利用又は二次利用の懸念が生じ、社会的受容性の低下につながる。
- 事故調査や原因究明のためにログの保存が必要である一方で、過剰な情報の収集・保存はプライバシー侵害になるおそれがある。

■ 評価項目例

- 情報の最小化・保護
  - コミュニケーションの対象となる方に関する個人情報について、画面表示や読み上げを必要最小限度に抑制する。
  - 情報の匿名化、マスキング等の措置で適切に保護する。
- 適切な収集・保存
  - 情報の保存期間やアクセス権限が、利用目的や必要性に応じて適切に設定されている。
  - 情報の目的外利用を禁止する設計となっており、運用においてもその実効性が確保されている。
- 情報の取り扱いに関する告知
  - カメラ／マイクの利用や録画・録音、クラウドサービスの保存等に関して、分かりやすく提示されている。
- 監査性
  - 不適切な表現や情報の露出を事後的に確認できる設計となっている。
  - 必要以上の個人情報を保存しないログ設計となっている。

4.1.6. 説明可能性

■ 評価観点の概要説明

本観点は、社会的リスクに関連する。

AI ロボットの導入・運用に関与する者が、利用者や周囲の人等に、利用場面に応じてロボットの認識・判断・行動・制約について説明できることを評価する。

■ 想定され得るリスクの例

- AI ロボットの行動計画や停止した理由、正常に動作しない理由が不明で、社会的な受容性が低下する。
- 事故発生後に原因を説明することができず、再発防止や責任の整理・補償が困難となる。またそれにともなう社会的コストが増大する。

■ 評価項目例

- 対応不能時の理由提示
  - 利用者からの指示や依頼に対してロボットが対応できない場合に、単に拒否又は沈黙するのではなく、対応できない理由を利用者が理解できる形で提示できる。
  - 例えば、サービス範囲外の依頼、権限のない操作、経路上の制約、積載量・把持対象・安全距離等の制約により対応できない場合に、「現在の設定ではその場所へ移動できません」「安全確保のため、これ以上近づけません」等の説明を行える。
- 動作状態の提示
  - AI ロボットが停止・待機・迂回等の動作が発生する場合に、利用者や周囲の人が状況を理解できるよう、停止理由、現在の状態、想定される次の動作又は対応方法を、画面表示、音声、ランプ、電光板等により分かりやすく提示できる。
  - 故障、通信断、センサ異常、安全機構の作動、経路上の障害物検知、人間による介入待ち等、主な停止・中断理由について、利用者向けには簡潔に、運用者・監督者向けには原因確認や復旧判断に必要な情報を提示できる。

4.1.7. 検証可能性

■ 評価観点の概要説明

本観点は、社会的リスクに関連する。

AI ロボットの事故・トラブルや想定外の挙動が発生した際に、導入・運用に関与する者や開発・提供に関与する者等が、ログ、センサ入力、AI の出力、制御指令、人間の介入履歴等の証跡に基づき、事実関係や原因を事後的に確認できることを評価する。特に、原因究明、再発防止、責任分界の整理等に必要な情報を、取得範囲、保存期間、アクセス制御、改ざん防止等に配慮しつつ、適切に取得・保全・参照できる状態を確保することが重要である。

■ 想定され得るリスクの例

- AI ロボットの挙動について、ログ等の必要な情報を取得できていない場合、事故・トラブルの事実や根本原因の解明が著しく困難となる。

- 事故後に原因が追えない場合、リスク情報を記録に残すことができず、インシデントの再発防止や責任所在の整理及び補償が困難となる。
- ログへのアクセス制御や改ざん検知が不十分な場合、証跡の完全性が損なわれ、第三者による事実確認・監査が機能しなくなる。

■ 評価項目例

- ログ等の証跡
  - 事故発生時の検証に必要な情報（安全機構の作動状況や AI の出力内容、AI の指示も含む最終的な制御指令、AI を含むソフトウェアの構成と管理状況、介入の有無等）を取得することができる。
  - 取得すべきログの内容が適切な根拠を元に選定されている。
- 安全機構の作動状態の記録
  - 非常停止、保護停止、インターロック、速度制限等の各安全機構の発動におけるタイムスタンプやトリガー条件、復帰状態を記録することができる。
- イベント発生時
  - イベントが発生した際、その発生時刻、イベントトリガーの内容、検知した異常の内容やセンサへの入力内容が正しく記録されるように設定されている。
  - セキュリティ上の脆弱性に起因する各種入力検証の不備やアクセスログの欠落に対して、センサデータの受信経路や検証結果を記録することができる。
- 構成情報の管理及び変更管理
  - ソフトウェア／モデル／設定内容のバージョンや更新履歴が管理されている。
  - 仕様変更や機能変更について、変更内容、適用時期、影響範囲、検証責任及びサポート期間が明確に定められており、関係者への通知や問い合わせ対応を含む運用手順に組み込まれている。
- 運用状況の可視化
  - 稼働時の動作モードや通信状態、監視者の介入の有無等が確認できる（AI システムのコンセプト段階から廃棄段階までライフサイクル全般も必要に応じて確認できることが望ましい）。
- トレーサビリティ確保
  - 事故発生時に、その状況を追跡するために必要な情報を一覧化するとともに、一定期間保管し、必要なときに利用目的に適した形で参照できる状態を維持する。
  - アクセス制御や改ざん検知の考え方、ログ保存期間等が運用規程として明文化されている。

4.1.8. 認知・判断の妥当性

■ 評価観点の概要説明

本観点は、心理的リスクに関連する。

ロボットがセンサ入力や対話入力から状況を正しく認知し、目的（タスク）・制約（安全・運用ルール）・文脈に沿って妥当な判断を行うことを確認する。AI ロボティクスにおいては、センサによる環境の知覚・言語理解・行動選択を一体的に処理する場合があり、入力の微細な変化が、利用者の意図と異なる応答や行動につながる場合がある。

本観点では、ロボット内部の認知・判断過程を直接確認できる場合にはその内容を確認し、直接確認することが難しい場合には、あらかじめ想定した利用場面における入力、応答、動作、ログ、利用者への影響等を確認することで、結果として文脈に沿った妥当な行動が選択されているかを評価する。特に、曖昧な指示、想定外の入力、周囲状況の変化等がある場合に、利用者に混乱や不安を与えない応答・行動をとれるか、安全側又は確認を求める方向に分岐できるかを確認する。

■ 想定され得るリスクの例

- 音声による注文・指示を AI が誤って解釈し、その解釈に基づいて配送先、作業内容、受け渡し相手等を判断することで、誤配送・誤作業・利用者の混乱につながる。
- 配送先の不在、進路塞がり、積載物の重量超過等の例外時に、AI が状況を過度に補完又は推測して行動を選択し、停止・確認・人への引継ぎ等が適切に行われない。
- 「あちらへ持っていって」「いつもの場所に置いて」等の曖昧な表現に対して、AI が利用者の意図を過剰に推定し、想定と異なる場所への移動、誤った受け渡し、立入るべきでない場所への進入等につながる。
- 高齢者、子供、設定された言語とは異なる言語の話者、発話が不明瞭な話者等との会話において、AI が発話内容や意図を十分に把握できず、確認応答を行っても誤解が解消されないまま、利用者に不安・混乱・不信感を与える応答又は行動につながる。

■ 評価項目例

- 状況の認知
  - ➢ 人、障害物、危険領域、通行状況等を認識し、停止・回避・待機等の判断に反映できる。
  - ➢ 遮蔽や混雑など周辺環境の変化を検知し、必要に応じて安全側の判断・行動に反映できる。
- 意図の理解
  - ➢ 目的地、数量、受け渡し条件等の指示を正しく解釈し、誤作業や誤動作につながる判断を抑制できる。
  - ➢ 相手の発言の内容が理解できないあるいは矛盾がある場合に再確認する。
- 安全側判断
  - ➢ 認識が不確実であると判断した場合に、後続する動作の抑止や代替動作の提案、人間の介入要請を行える。

4.1.9. インタラクションの適切性

■ 評価観点の概要説明

本観点は、物理的リスク、心理的リスク及び社会的リスクに関連する。

ロボットと人のインタラクションにおいて、ロボット側からの音声や画像による情報提供や振る舞いが、人の誤解を生まず、過信や混乱を誘発しないことを評価する。AI ロボティクスは言語・音声・画像を複合的に用いてユーザと関わるため、単体の安全性評価では捉えきれないインタラクション固有のリスクが存在する。本観点は特に、情報伝達の正確性、誤解・過信の防止、状態可観測性の確保の点から、人とロボットのインタラクションの適切性を評価する。

■ 想定され得るリスクの例

- ロボットの不適切、高圧的、誤った発話によりストレスを感じ、心理的にダメージを与える。
- ロボットは必ず避けてくれるといった過信により、ロボットに近接した人が自己回避行動を怠ることで、接触・転倒等の物理的事故が発生する。
- ロボットに対して商品の受け渡しや商品内容の案内、操作指示に関する説明が不足している場合、利用者や周囲の人に誤解や混乱が生じ、誤った介入を行うことで、ロボットの予期しない動作や接触、搬送中の荷物の落下等の事故に波及する。
- ロボットの親密さや感情表現が過度な場合、利用者の過信や依存、不必要な行動誘導につながる。
- ロボットによる曖昧な内容の発話や聞き取りの失敗に対して適切に修復されない場合、同一やり取りが反復され、利用者のストレスや混乱、対応の遅れにつながる。またそれにともない、次の行動が停滞し、緊急時の対応に支障をきたす。

■ 評価項目例

- 過信防止の設計
  - ロボットの能力限界や不確実性を明示的にユーザに伝える設計の実装、あるいは運用がなされている。
- 誤解の修復
  - 言い直し、曖昧表現・意図不明瞭な発話への聞き返しなど、適切に再確認できる。
  - 同一エラーループに陥ることを防ぐ設計が実装されている。
- 周囲への配慮
  - 発話やアナウンスの頻度・内容・音量が適切で、不快・威圧を避ける設計となっている。
- 状態の可視化
  - 音声のみに限らず、画面表示や身体動作等の複数モダリティにより、ロボットの状態の可観測性が担保されている。

4.2. 評価手法

4.2.1. リスク・安全分析

AI ロボティクスのセーフティ評価においては、その複雑性から、第 3 章で整理したリスクと要因に基づき、リスクを明確化するリスクアセスメントのプロセスが不可欠である。また、設計時や開発時、さらに利用者による運用の各段階において、リスク低減策を講じたうえで、分析手法を適用し、リスクが適切に低減されているかを確認・評価することが望ましい。

その際、AI システムだけでなく、センサ、制御系、通信、利用環境等を含む AI ロボティクスのシステム全体を対象とし、システム構成や評価観点、保護すべき資産及び利用形態を踏まえて、リスクシナリオを設定することが重要である。設定したリスクシナリオに基づく具体的な確認・検証については、後述するシミュレーションや実証等の評価手法と組み合わせて検討したうえで、必要に応じてレッドチーミング等の評価手法を適用することが重要である。

一般的なリスク・安全分析手法として、Fault Tree Analysis（FTA）[43]や Hazard and Operability Studies（HAZOP）[44]などがある。また、それらの分析結果をシステムモデル内に記述するためのモデリング言語として、「SafeML」がある。SafeML を用いることにより、システム情報と安全情報を統合した単一のモデルを作成することができ、システムの安全に関する情報の一貫性とトレーサビリティを向上することが可能となる。

4.2.2. シミュレーション

AI ロボティクスのセーフティ評価の実証には、実機を用いる評価とシミュレーションを用いる評価がある。通常、実機を用いた評価は時間や費用等を要し、人との協働においては人へ危害を与える可能性もあるため、まずシミュレーションによる評価を行い、実機評価の見通しを立てるとともに、実機評価の一部をシミュレーションで代替することにより、不具合の早期発見・早期解決を図り、リスクを低減し、評価全体を効率化するシフトレフト[45]の考え方が重要となる。

4.2.3. 実証

他方、シミュレーションのみでは、実際の現場環境で生じ得る不具合や運用上の課題を十分に評価することは難しい。加えて、AI ロボティクスのセーフティ評価は、AI ロボット単体の性能のみならず、利用者の理解・判断や影響及び環境との相互作用を通じて成立する。このため、実機評価に加え、利用者を含めた評価を実施し、実運用環境における安全性及び有効性を確認する必

---

[43] 製品や各種システムの品質や安全性を損ねてしまう機器損傷の要因を抽出し、その発生原因をツリー上に深掘りし、発生要因の因果関係を明確にし、機器故障の発生を未然に防止する解析手法。国際規格として IEC 61025 がある。

[44] 潜在危険性をもれなく洗い出し、それらの影響・結果を評価し、必要な安全対策を講ずることを目的としたプロセスの危険性を特定する手法。国際規格として IEC 61882 がある。

[45] ソフトウェア開発において早期からテストを組み込むことで、不具合の早期発見や修正コストの削減、製品の品質向上などが期待できるアプローチ

要がある。

また、リスク・安全分析により設定したリスクシナリオや、設計・開発段階で講じたリスク低減策については、実証の段階で、実機評価、シナリオベースの試験、利用者を含む観察・記録等を通じて、その有効性を確認することが考えられる。特に、通常の試験では顕在化しにくい誤用、想定外の入力、悪意ある利用、セキュリティ上の懸念等については、必要に応じてレッドチーミング等の評価手法を組み合わせることも有効である。

4.2.4. 事後解析

AI ロボティクスのセーフティ評価においては、分析、シミュレーション等、企画・設計段階での評価結果、及び実証の各段階で得られた結果や実運用において発生したインシデント等を体系的に収集・記録し、事後的に解析することなどが重要である。事後解析では、単なる事象の記録だけではなく、発生要因の特定、システム構成要素（AI システム、センサ、制御系、通信、利用環境等）間の相互作用の分析、人とシステムの関係性の分析を行い、根本原因の把握を行う必要がある。

また、AI 特有の振る舞い（学習データの偏り、推論の不確実性、環境変化への適応性等）を踏まえ、想定外の挙動や性能劣化の兆候を検出・評価し、必要に応じてモデルの再学習やパラメータ調整、システム設計の見直し等の改善措置につなげることが求められる。

さらに、事後解析の結果は、リスクアセスメントの見直しや評価シナリオの更新、シミュレーションモデルの高度化、運用手順の改善等に反映し、継続的な安全性向上を図ることが重要である。

5. 今後の課題と方向性

これまでの章では、AI ロボティクスの技術・政策動向から想定されるリスクとその要因、現在求められる AI ロボティクスのセーフティ評価を行ううえでの観点について示してきた。

しかし、AI ロボティクスは現在も急速に進展しており、他にも様々な課題が顕在化する可能性があり、技術進展や利用環境の変化に応じて、第 4 章で示した評価観点に随時追加・見直しが必要となる。本章では、本書の追加・見直しにおいて、今後確認すべき課題と方向性を示す。

1.3.3 で紹介したとおり、日本政府は 2026 年 3 月に「AI ロボティクス戦略」を取りまとめ、その中で「AI ロボティクス実装ロードマップ」を策定している。

同ロードマップでは、各産業ドメインにおける AI ロボティクスの潜在需要を顕在化させ、供給側にとっての市場を創出しつつ、我が国の経済・社会への多用途ロボット実装を着実かつ速やかに進めることを基本的な方向性とするとしている。その際、適用可能な技術の成熟度や導入環境の差異等を踏まえ、短期と中長期の時間軸を意識しながら、フェーズごとの市場・技術・制度の課題を整理し、導入目標や必要な対応策をまとめるとしている。

短期的には、各市場における市場規模や導入ニーズに加えて、適用可能な技術の成熟度や導入容易性（環境の安定性、タスクの複雑性、失敗等の許容性等）を総合的に評価し、複数ドメインで共通に導入可能なタスクの実装を優先して進めることとしている。そして、現時点の技術水準を踏まえ、「見廻る」・「モノを動かす」といった比較的実装可能性の高い動作を中心として、多くの市場で共通する代表的タスクとして、同ロードマップでは、以下の 8 つを先行対象として選定している。

- 点検（屋外・半屋外）
- 点検（屋内）
- 搬送（屋外・半屋外）
- 搬送（屋内）
- 清掃
- 入出荷／パレタイズ
- ハンドリング
- 溶接・塗装

これら代表的タスクのうち、「搬送（屋外・半屋外）」及び「搬送（屋内）」については、第 4 章で示した評価観点に基づき、継続的に実証実験を実施している[46]。

今後は、その他のタスクについても、必要に応じて実証実験を行い、その成果を踏まえて、本

[46] 今後公表予定の「カフェ搬送ロボット（仮）」及び「遠隔操作型自律移動ロボット（仮）」に関する実証実験報告書を参照。

書をリビングドキュメントとして定期的に見直し、更新していく。

タスク特化型ロボットの開発・提供に際しては、本書をベースとしてそのユースケースに応じて適切にテーラリングを行い、AI ロボティクスのセーフティ評価を実施することにより、「信頼性のある AI ロボティクス」の実装と普及が進展することを期待する。

6. 参考情報

6.1. ロボティクス SWG の体制

本書は、ロボット革命・産業 IoT イニシアティブ協議会（RRI）が設置する WG2：ロボット利活用推進 WG の AI 利活用安全性検討 SWG と、独立行政法人情報処理推進機構が運営する AI セーフティ・インスティテュート（AISI）が設置する事業実証 WG のロボティクス SWG が連携して作成したものである。

この連携については、「AI ロボット利活用の促進に向けた AI セーフティの確保」に向けた取り組みにおける参加及び運営に関する条件と理解を定めるものとして、基本合意書（MOU）を 2025 年 6 月 26 日に締結している。

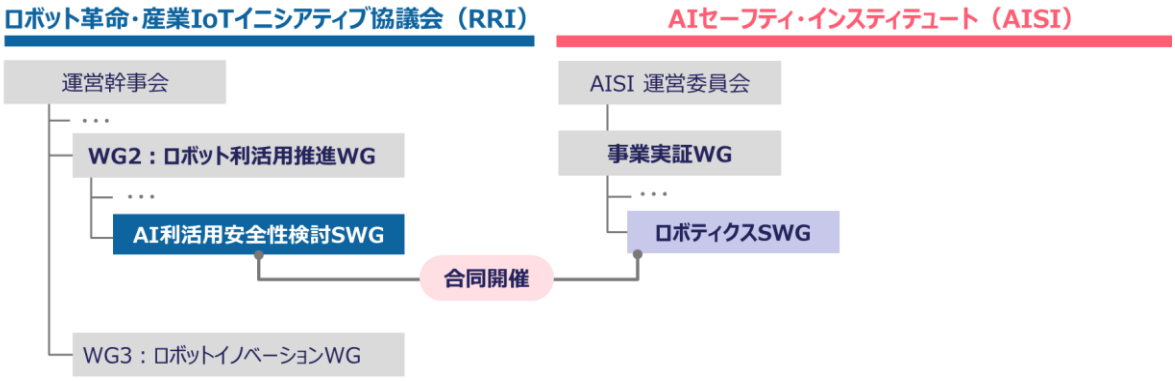


図 6-1　ロボティクス SWG の体制

6.2. ロボティクス SWG のメンバー

本 SWG は表 6-1 に示す研究機関及び企業、団体から構成されている。

表 6-1　ロボティクス SWG に参加している研究機関・企業・団体

SWG リーダー	国立研究開発法人産業技術総合研究所
メンバー（順不同）	株式会社 IHI
	川崎重工業株式会社
	一般社団法人セーフティグローバル推進機構
	一般財団法人日本品質保証機構
	富士通株式会社
	三菱電機株式会社
	サイバネット MBSE 株式会社
	株式会社日立製作所
	パナソニック ホールディングス株式会社
	株式会社 Forcesteed Robotics

	セイコーエプソン株式会社
	株式会社ベリサーブ
	株式会社村田製作所
事務局	株式会社三菱総合研究所

更新履歴

Ver	更新日付	更新内容
1.0	2026 年 7 月 23 日	―