

AIロボティクスに関するセーフティ評価観点ガイド (概要版)

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

2026年7月23日

AISI Japan
AI Safety Institute

- 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提供。
- 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。

ガイドの構成

1. 背景・目的

- ◆ 背景、目的、本書のスコープ、想定するユースケース、対象読者、本書および別冊の使い方、本書で使用する用語

2. 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・利活用事例、AIセーフティを含む政策・規制・標準化動向

3. AIロボットのリスクと要因

- ◆ 想定されるリスク、リスク発生要因、セーフティ評価の必要性

4. AIロボティクスに関するセーフティ評価観点

- ◆ 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

5. 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た課題や方向性

- ロボティクス分野におけるAIの利用に伴い、事業者がAIセーフティに配慮したAIロボットの開発・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。
- 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。ユースケースは「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」を想定。
- 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が想定読者。

本ガイドの Scope

対象読者	AIロボティクスの供給側（特に開発や提供・運用に直接携わるステークホルダー）	
対象ロボット	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	
対象のAI	テキストベースの大規模言語モデル（LLM）、画像認識AI、音声認識AI （次版以降、Vision Language Action（VLA）やロボット基盤モデル等を検討）	
対象のユースケース	カテゴリ	ユースケース例
	飲食店等における商品搬送ロボット	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。
	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。

- AIロボティクスに関する技術動向と利活用事例を概観したうえで、各国の政策・業界団体等によるAIセーフティへの取組、およびAIロボティクスの政策・業界動向を紹介。
- AIの信頼性とリスク、AIシステムと機能安全の国際標準化の動向についても紹介。

AIロボティクスの技術動向

- ・ **主な技術動向**
 - 認識・判断・行動を統合したAIロボティクスへ発展し、マルチモーダルAIの活用により、人と柔軟に協調する新たなロボットの利用形態が実現されつつある。
 - ロボット基盤モデルやVLA（Vision-Language-Action）モデルの発展により、多様なタスクへの汎化や未知環境への適応、自律的な行動が期待されている。
 - センサや計算機などハードウェアの進展との融合により、適用分野は製造業から物流、サービス、ヘルスケア等へ。
 - 将来的には、世界モデルやフィジカルAIを活用し、現実世界を理解して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。

AIロボティクスの政策・業界動向

- ・ **各国動向**
 - 日本：人工知能基本計画（2025）、AIロボティクス戦略（2026）
 - 米国：A Roadmap for US Robotics: Robotics for a Better Tomorrow（2024）
 - 欧州：Horizon Europe
 - 中国：AI Safety Governance Framework 2.0（2025）
- ・ **業界動向**
 - 日本：JARA「ロボット産業ビジョン2050」、AIRoA設立
 - 米国：A3「ホワイトペーパー AI in Automation」
 - 英国：Adra「Policy Paper and Technology Roadmap」

- AIロボットに関する想定リスクを3つの類型、リスク要因を4つに分類。
- 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を中心とする観点に加え、心理的リスクおよび社会的リスクも含めて、総合的に評価することが必要。

AIロボットのリスク類型

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）

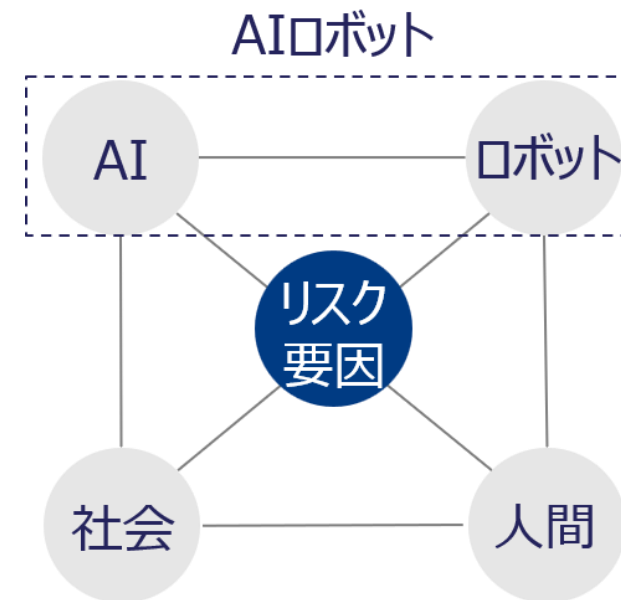
社会的リスク



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

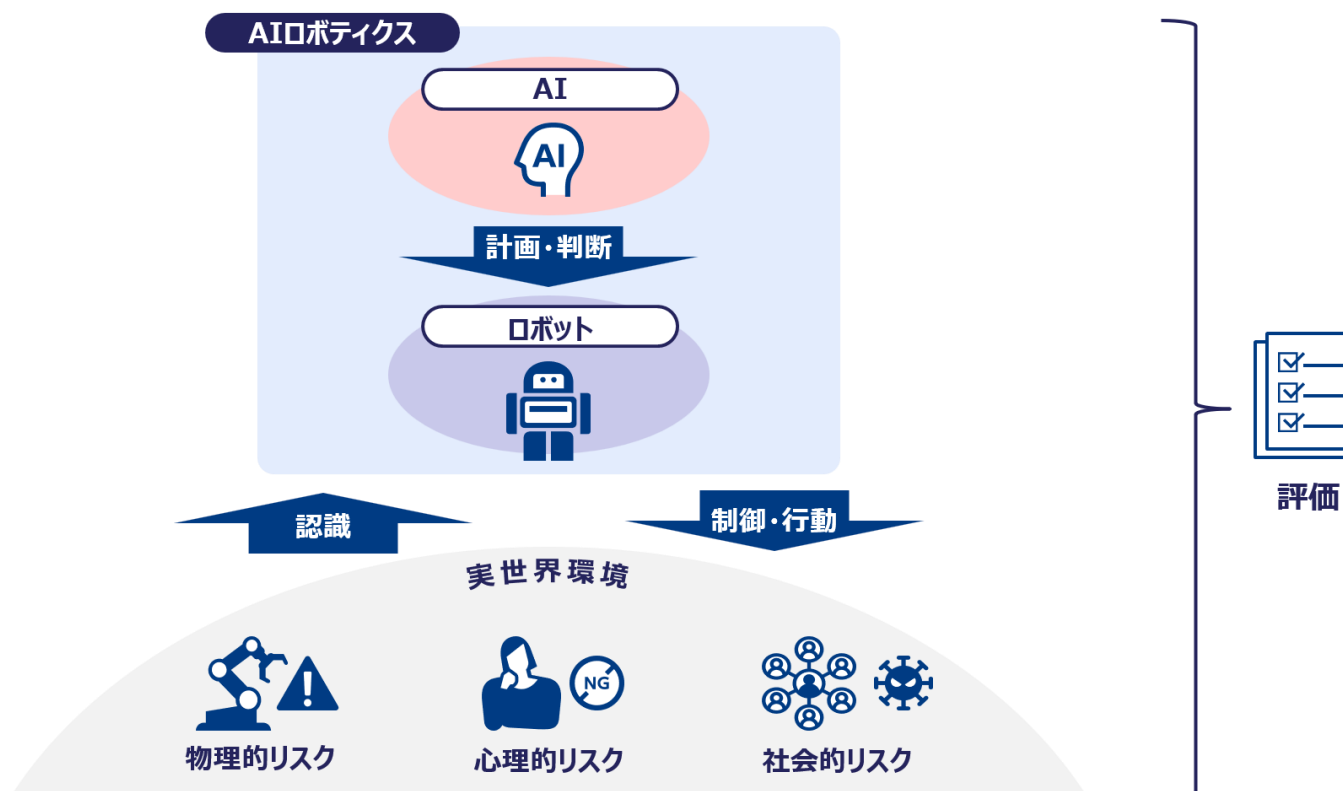
AIロボットのリスク要因

「AI」、「ロボット」に加えて、AIおよびロボットを扱う「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」が考えられる。



第4章：AIロボティクスに関するセーフティ評価観点

- AIロボティクスは、周囲の環境や対象を捉える「認識」、行す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。



第4章：AIロボティクスに関するセーフティ評価観点

- 主要なリスクについて評価を行う際に必要となる評価観点を整理。ただし、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。

評価観点	関連リスク			評価観点の概要
	物理的	心理的	社会的	
物理的安全性	●			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。
運用合理性	●	●		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。
安定性	●	●		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。
人間の適切な介在	●	●	●	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。
プライバシー保護			●	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。
説明可能性			●	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。
検証可能性			●	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。
認知・判断の妥当性		●		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。
インタラクションの適切性	●	●	●	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ロボティクスのセーフティ評価に不可欠な評価手法と評価を実施する上でのポイントを整理。

評価手法

実施する上でのポイント

リスク・安全分析

- AIシステムだけでなくセンサ・制御系・通信・利用環境等を含むシステム全体を対象に、システム構成・評価観点・保護すべき資産・利用形態を踏まえてリスクシナリオを設定する。
- 確認・検証はシミュレーションや実証等と組み合わせ、必要に応じてレッドチーミング等も適用する。

シミュレーション

- 実機評価は時間・費用を要し、人との協働では危害のおそれもある。まずシミュレーションで評価の見通しを立て、実機評価の一部を代替することで、不具合の早期発見・早期解決とリスク低減、評価全体の効率化を図る「シフトレフト」の考え方が重要となる。

実証

- 実機評価に加え、利用者を含めた評価を行い、実運用環境での安全性・有効性を確認する。
- 設定したリスクシナリオや設計・開発段階のリスク低減策は、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。

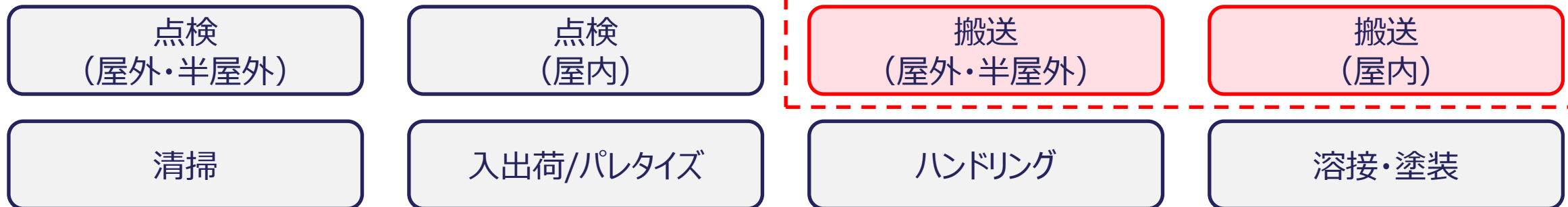
事後解析

- インシデントが発生した場合は、事象の記録にとどまらず、発生要因の特定や構成要素（AIシステム・センサ・制御系・通信・利用環境等）間の相互作用、人とシステムの関係性を分析し、根本原因を把握する。
- 事後解析の結果を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。

- 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。
- 2026年3月に策定された「AIロボティクス実装ロードマップ」の短期的に優先して実装に取り組むべきタスクのうち、「搬送（屋外・半屋外）」および「搬送（屋内）」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する。
- 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、本書をリビングドキュメントとして定期的に見直し、更新する。

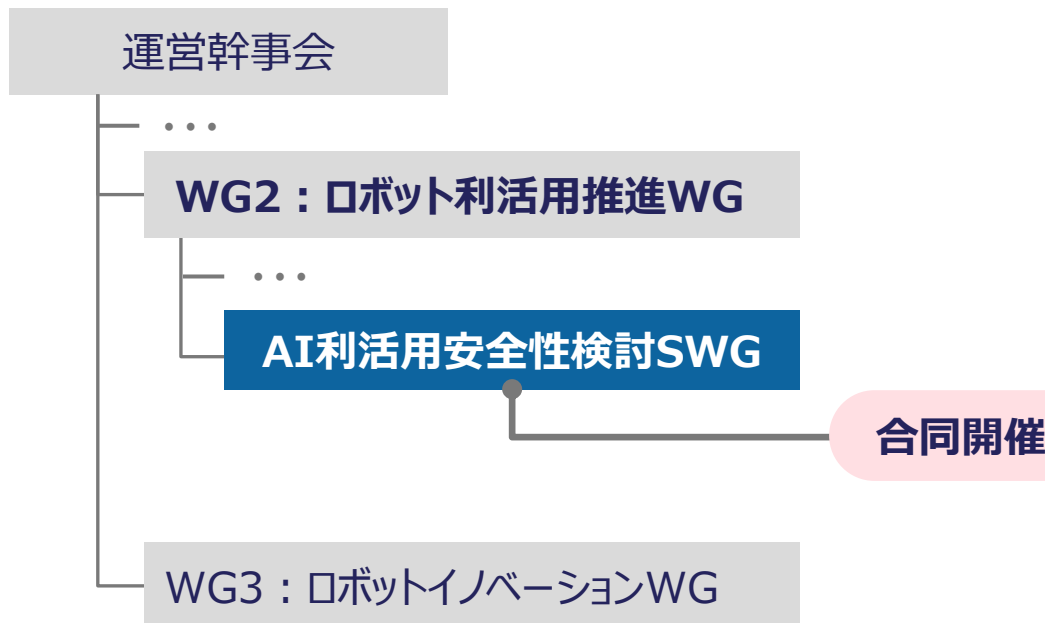
AIロボティクスのセーフティ評価により、「信頼性のあるAIロボティクス」の実装と普及の進展へ

□ AIロボティクス実装ロードマップ 優先的に実装に取り組む共通タスク



(参考) ロボティクスSWG体制・参加組織

ロボット革命・産業IoTイニシアティブ協議会（RRI）



AIセーフティ・インスティテュート（AISI）



合同開催

国研）産業技術総合研究所（SWGリーダー）
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Forcesteel 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
株式会社三菱総合研究所（事務局）

AISI

Japan AI Safety Institute