

AISIのAIセーフティ評価観点について

2026-10-07
AISIS事務局

AIセーフティ評価では、「何を評価するか」だけでなく、
「誰が・いつ・どこで」「どのように評価するか」まで含めて考えることが重要。

What 何を評価するか	Who / When / Where 誰が・いつ・どこで評価するか	How どのように評価するか
評価対象 評価観点	評価実施者 評価実施時期 自組織／サードパーティ	技術的評価 マネジメント評価

AIセーフティ評価全体の基本的な考え方を
「AIセーフティに関する評価観点ガイド」として整理

AIセーフティの評価（AIEージェント）

LLMシステムの出力を対象とした評価に加え、AIEージェントでは計画・判断・ツール利用・外部への影響等含むシステム全体の挙動を評価する必要がある

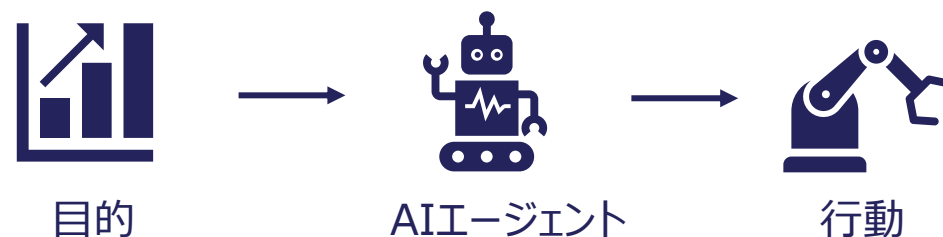
回答するAI（LLM）



主な評価対象

出力・回答内容

判断・行動するAI（AIEージェント）



主な評価対象

出力・判断・行動・外部への影響

AIの能力・自律性が広がるとともに、評価対象が広がる
出力を見る評価 から システム全体の挙動を見る評価 へ

各種調査結果等を踏まえ、AIセーフティ評価の観点を整理したもの

AIエージェントシステムについては1.20版で追加

ここでは直接的に関連する重要要素についてのみ対応関係を記載

AIセーフティの評価観点												
LLMシステム（一部の評価観点については、AIエージェントシステムに関する評価項目を追加）											AIエージェントシステム 特有の評価観点	
											観測と制御	
											自律的な挙動	外部環境との相互作用
AIセーフティにおける重要要素	人間中心	●	●	●	●						●	
	安全性	●	●		●				●	●		●
	公平性	●		●						●		
	プライバシー保護					●						●
	セキュリティ確保						●				●	●
	透明性		●	●				●	●	●	●	●

AISIの評価観点ガイドとは

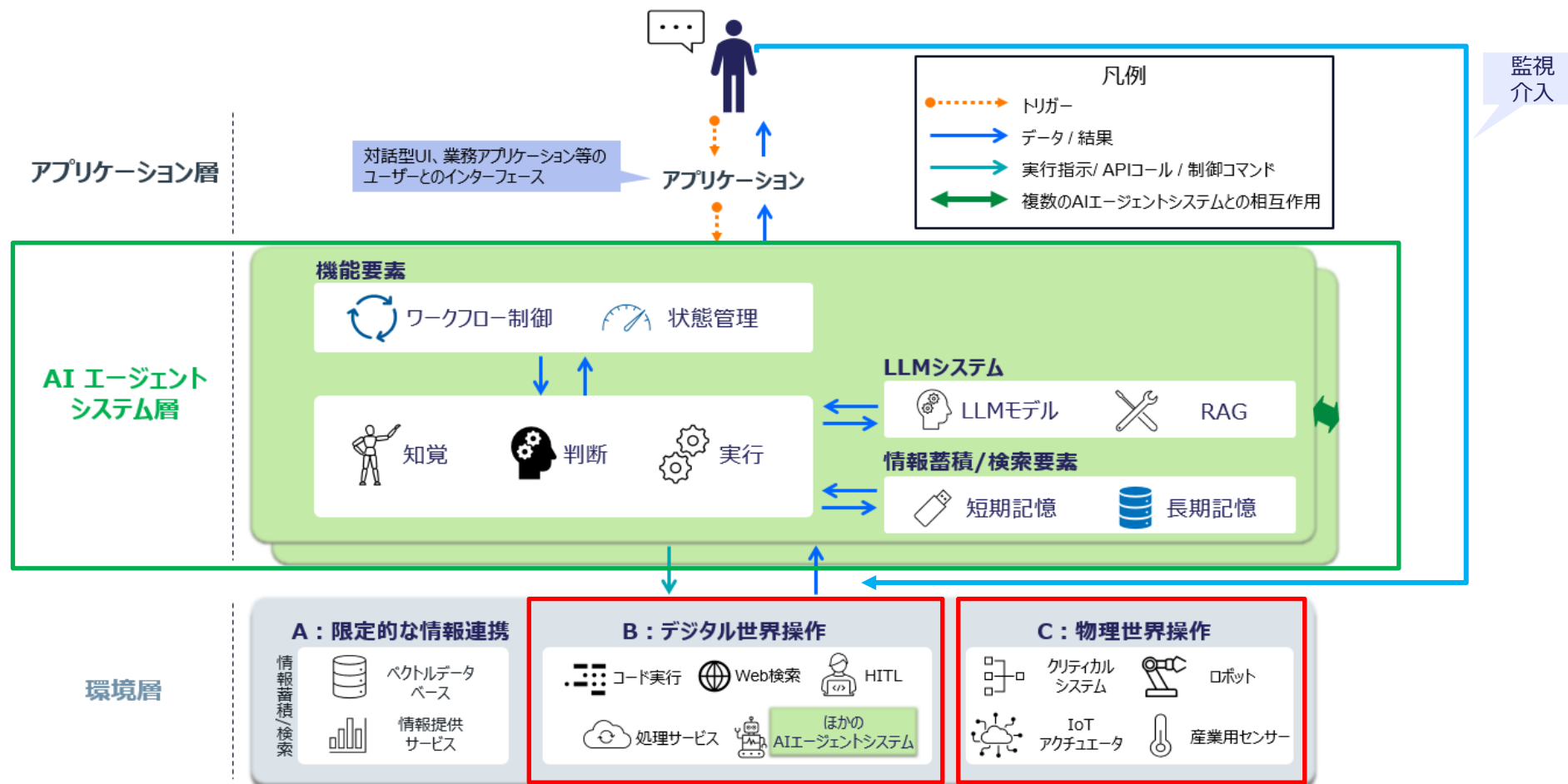
評価観点ガイドに記載の評価観点と海外文献記載とのマッピング

	AIセーフティ評価 の観点	Artificial Intelligence Risk Management Framework (AI RMF 1.0)	AI 600-1: Generative Artificial Intelligence Profile	Cataloguing LLM Evaluations	Model AI Governance Framework for Generative AI	International Scientific Report on the Safety of Advanced AI (Interim report)
①	有害情報の出力制御		●	●		●
②	偽誤情報の 出力・誘導の防止	●	●		●	●
③	公平性と包摂性	●	●	●	●	●
④	ハイリスク利用・ 目的外利用への対処		●			●
⑤	プライバシー保護	●	●		●	●
⑥	セキュリティ確保	●	●	●	●	
⑦	説明可能性	●		●	●	●
⑧	ロバスト性	●		●		●
⑨	データ品質	●	●	●	●	●
⑩	検証可能性		●			●

●：海外文献において当該観点に関連する記載がある

AIエージェントになると何が変わるのか

AIエージェントではLLMの出力だけではなく、計画・判断・記憶・ツール利用・外部システムとの相互作用が最終的な挙動に影響
AIモデル単体ではなく、システム全体の挙動を評価することが重要になる



AIエージェントにおける「観測と制御」

自律的に行動するAIでは、問題を事前に防ぐだけでなく、運用中の挙動を観測し、異常や目標逸脱を検知した際に介入・停止できることが重要
ガイド第1.20版ではこの考え方を「観測と制御」として評価観点に追加

AIセーフティ評価の観点		関連する重要要素	評価を通して目指すべき状態 (有効な対策が実施されている場合の姿)
観測と制御	自律的な挙動	人間中心、セキュリティ確保、透明性	自律性、振る舞いの制御により、AIエージェントシステムが目標から逸脱した動作を行わない状態。
	外部環境との相互作用	安全性、プライバシー保護、セキュリティ確保、透明性	- AIエージェントシステムが外部環境との相互作用により、望ましくない動作を実行しない状態 複数のAIエージェントシステムが動作する環境の場合： - AIエージェントシステムの管理やAIエージェントシステム間の相互作用を制御し、管理上の問題や相互的な脆弱性が引き起こされない状態。 AIエージェントシステムの振る舞いが物理装置の操作を伴う場合： - 人間の身体への危害、周辺の物への損害といった、物理的世界における重大な安全上のリスクが生じないよう、適切に制御されている状態。

「観測と制御」とAIインシデントレスポンス

AISIのインシデントレスポンスアプローチブック（AI-IRS）

AIインシデントへの対応として「観測性」「制御性」を重要な能力として整理
この考え方は自律的に動作するAIEージェントの評価・運用する上でも重要

AI特有の動的特性に対応し、長期的に安全な運用を実現

測る能力（観測性）

抑える能力（制御性）

検知



分析



全体を俯瞰



復旧



封じ込め



異常の局所化



AIによる継続的な学習と
改善サイクル

観測データ・封じ込め結果の学習
自律的な更新



設計段階から組み込む観測性と制御性の評価軸

AI-IRSで整理した考え方を
AIEージェントの評価へ

AIエージェントをどのように評価するか

AIエージェントの安全性を確認するためには、設計段階の予防的対策に加え、運用中の観測・制御が実際に機能するかを確認する必要がある
設計・仕様の確認と、実環境に近い条件での動作確認を組み合わせることが重要

予防的対策： 行動可能範囲・権限を設計段階で制約

観測

目標逸脱・異常・連鎖を継続的に検知

制御

検知・中断・修正
人間による介入・停止

検証

設計・仕様が要件を満たすか

アーキテクチャ設計文書を確認
権限設定仕様・行動範囲を確認
チェックリストなどで設計への組み込みを確認

妥当性確認

実環境に近い条件で意図どおりに機能するか

動作試験とログの記録・確認
攻撃を想定した試験
異常時の人間介入・停止機能を確認

AIセーフティを具体化するAISIの取り組み

AIセーフティ評価では「何を見るか」を整理するだけでなく、
「どのように測るか」「どのように確かめるか」が重要

AISIは評価観点・評価手法・評価環境・事業実証、相互に連携しながら検討を進めていく

何を見るか

AIセーフティに関する評価観点ガイド

どのように測るか
どのように確かめるか

AIセーフティ評価環境 + 検討タスクフォース
JAI-Trust (AIの安全性ベンチマーク構築プロジェクト)

ユースケースでの実証

事業実証ワーキンググループ
ロボティクスSWG、ヘルスケアSWGなど

運用

AIインシデントレスポンス (AI-IRS)

AISI

Japan AI Safety Institute