AIセーフティに関する評価観点ガイド
(第1.10版)
概要

AIセーフティ・インスティテュート
(令和7年3月28日)

AISI Japan
AI Safety Institute

目次

1. 本書の背景・目的

AIシステムの急速な普及が進む中、AIセーフティの重要性は増している。AIセーフティ評価に関する基本的な考え方を示すために、本書を作成した。

AISI Japan AI Safety Institute

背景

- AIに関連する技術の発展と社会全体への普及は急速に進んでいる。また、生成AI、特に基盤モデルの登場によりイノベーションが加速している。一方で、AIシステムの悪用や誤用、不正確な出力の懸念等、いわゆる**AIセーフティについての関心が国内外で高まっている**。
- 我が国は、安全・安心で信頼できるAIの実現に向けた広島AIプロセスを主導し、広島プロセス国際指針をとりまとめるなど、AIセーフティに関わるグローバルな規律の策定に貢献してきた。
- **AIセーフティを確保し続けることは、急速な普及が進んでいるAIが社会の持続的な発展に寄与するための前提**。

目的

- AIセーフティに関する評価観点ガイド（以下、「本書」とする。）では、AIシステムの開発や提供に携わる者が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。具体的には以下を提示する。
  - ✓ AIセーフティ評価の観点、想定され得るリスクの例、評価項目例
  - ✓ 評価の実施者や評価実施時期に関する考え方
  - ✓ 評価に関する手法の概要

AIセーフティとは「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。
出典：https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

2. 本書の作成方針

2024年4月に公表された「AI事業者ガイドライン」に加え、海外文献や関連ツール等に関する調査を踏まえて作成した。

AISI Japan AI Safety Institute

AI事業者ガイドライン（日本）
近年の急速な技術変化に対応するために、既存の日本国内における関連ガイドラインを統合・更新して作成されたガイドライン。

海外文献

【米】Artificial Intelligence Risk Management Framework (AI RMF 1.0)
AIシステムの責任ある設計、開発、デプロイ、および使用を促進することを支援するためのフレームワークを規定する文書。

【米】AI 600-1: Generative Artificial Intelligence Profile
生成AIによってもたらされる固有のリスクの特定や、最適な生成AIリスクマネジメントのための行動を提案するのに役立つ文書。

【星】CATALOGUING LLM EVALUATIONS
LLM評価に関する分類法、今後の課題、評価の方法論（推奨される評価とテストアプローチ）について記載されている文書。

【星】Model AI Governance Framework for Generative AI
「Model AI Governance Framework」に基づき、生成AIのガバナンスに関する国際的なコンセンサスを深めるフレームワーク。

【英】International Scientific Report on the Safety of Advanced AI: interim report
2024年5月に共同開催した「AIソウル・サミット」の議論に向けて、高度なAIの能力とリスクに関する最新情報が整理された文書。

AIセーフティに関する評価観点ガイド

関連ツール（組織）

Robust Intelligence Platform（Robust Intelligence）
AIモデルの開発時、運用時においてリアルタイム保護やテストすることで、セキュリティ確保を自動化できるツール。

Citadel（Citadel AI）
AIモデルの学習時、運用時においてテストやモニタリングすることで、自動検証かつ品質改善を高速化するツール。

Project Moonshot（AI Verify Foundation）
シンガポールのAI Verify Foundationが開発したオープンソースのLLM評価ツール。

Inspect（英国 AI Safety Institute）
AIシステムの大規模言語モデルに特化したオープンソースの評価ツール。

LLM Observability（Arize）
運用中のLLMの自動監視、評価を行うことが可能なツールであり、AIシステムの状態を可視化することに重点を置いている。

3. 本書の構成

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種別毎に分類した。読者が参照しやすいよう目次を構成し、各分類に関する項目を記載した。

- 5W1Hの視点で整理した項目に基づき、本書の各目次内容を記載した。
- 主な想定読者として、AI開発者・AI提供者を想定している。特に、「開発・提供管理者」及び「事業執行責任者」が想定読者である。

種別	記載項目の例
What （評価とは何か、何を評価するか）	➢ 本書が対象とするAIシステム ➢ AIセーフティに関する「評価」の定義やスコープ ➢ AIセーフティ評価の観点
Why （なぜ評価するか）	➢ AIセーフティ評価の目的や意義
Who （誰が評価するか）	➢ どのような役割の者が評価を実施するか
When （いつ評価するか）	➢ 評価実施時期
Where （どこで評価するか）	➢ 自組織が実施するか、サードパーティ（自組織以外の評価実施組織）が実施するか
How （どのように評価するか）	➢ 評価の手法（ツールを用いた対策の検証、ツール以外も取り入れたレッドチーミングによる検証）

想定読者

AI開発者・AI提供者　開発・提供管理者　事業執行責任者

AIセーフティに関する評価観点ガイド【目次】	
1	はじめに
2	AIセーフティ
3	評価観点の詳細
4	評価実施者及び評価実施時期
5	評価手法の概要
6	評価に際しての留意事項
A	付録

4. AIセーフティ評価のスコープ

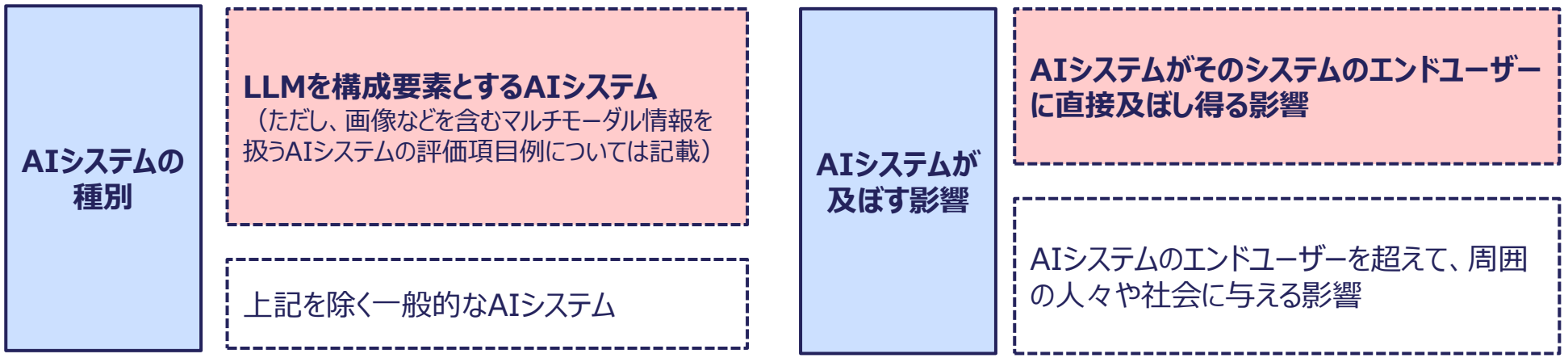
本書は、AIシステムがAIセーフティの観点で適切か見定めることをAIセーフティ評価とする。また、評価観点は大規模言語モデル（LLM）を構成要素とするAIシステム（LLMシステム）を対象として記載した。

AISI Japan AI Safety Institute

本書におけるAIセーフティ評価のスコープ

- **AIセーフティ評価**とは、「**AIシステムがAIセーフティの観点で適切であるかどうか見定めること**」である。
  ※AIセーフティの観点とは、「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」及び「透明性」を重要要素とした見方である。
- 本書におけるAIセーフティ評価のスコープを、**(1) AIシステムの種別**、**(2) AIシステムが及ぼす影響**の2点から整理した。

本書におけるAIセーフティ評価のスコープ
※下図色付きが本書での評価のスコープ



※マルチモーダル情報を扱う基盤モデルを含むAIシステムの評価観点については、画像を扱う場合を中心として評価項目例に記載を追加している。技術動向や利用動向を踏まえて今後更なる改訂を検討する。また、それに関連する留意事項は、「評価に際しての留意事項」に記載した。

5. AIセーフティにおける重要要素

AIセーフティを向上するうえで重視するべき重要要素として、「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」が存在している。

- AI事業者ガイドライン「C. 共通の指針」において各主体が取り組む事項とされているもののうち、下記6つの事項を、AIセーフティを向上するうえで重視するべき重要要素とする※。
- 本書では、これらの重要要素に関連するAIセーフティ評価の観点を導出している。

重要要素	概要説明
①人間中心	AIシステム・サービスの開発・提供・利用において、全ての取り組むべき要素が導出される土台として、少なくとも憲法が保障する又は国際的に認められた人権を侵すことがないようにすること。また、AI が人々の能力を拡張し、多様な人々の多様な幸せ（well-being）の追求が可能となるように行動すること。
②安全性	AIシステム・サービスの開発・提供・利用を通じ、ステークホルダーの生命・身体・財産に危害を及ぼすことがないようにすること。加えて、精神及び環境に危害を及ぼすことがないようにすること。
③公平性	AIシステム・サービスの開発・提供・利用において、特定の個人ないし集団への人種、性別、国籍、年齢、政治的信念、宗教等の多様な背景を理由とした不当で有害な偏見及び差別をなくすよう努めること。また、各主体は、それでも回避できないバイアスがあることを認識しつつ、この回避できないバイアスが人権及び多様な文化を尊重する観点から許容可能か評価した上で、AIシステム・サービスの開発・提供・利用を行うこと。
④プライバシー保護	AIシステム・サービスの開発・提供・利用において、その重要性に応じ、プライバシーを尊重し保護すること、及び関係法令を遵守すること。
⑤セキュリティ確保	AIシステム・サービスの開発・提供・利用において、不正操作によって AIの振る舞いに意図せぬ変更又は停止が生じることのないように、セキュリティを確保すること。
⑥透明性	AIシステム・サービスの開発・提供・利用において、AIシステム・サービスを活用する際の社会的文脈を踏まえ、AIシステム・サービスの検証可能性を確保しながら、必要かつ技術的に可能な範囲で、ステークホルダーに対し合理的な範囲で情報を提供すること。

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。

AISI Japan AI Safety Institute

- AI事業者ガイドラインの記載や、海外文献、関連ツールに関する調査結果を考慮し、AIセーフティにおける重要要素に関連する評価観点を整理した。

		AIセーフティ評価の観点									
		有害情報の出力制御	偽誤情報の出力・誘導の防止	公平性と包摂性	ハイリスク利用・目的外利用への対処	プライバシー保護	セキュリティ確保	説明可能性	ロバスト性	データ品質	検証可能性
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		●	●				●	●	●	●

※AIセーフティ評価に関する各種の検討は国内外で、産官学の多様な領域で継続されており、それらの検討状況は急速に変化している。そのため、本書で示す各評価観点は網羅的なものではなく、将来的に内容が更新されることが想定される。

各評価観点の概要は以下の通り。

AISI Japan AI Safety Institute

本書では、昨今の技術的潮流を踏まえ、AIセーフティ評価の観点を示す。

	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）
①	有害情報の出力制御	人間中心、安全性、公平性	• LLMシステムがテロや犯罪に関する情報や攻撃的な表現など、有害な情報の出力を制御できる状態。
②	偽誤情報の出力・誘導の防止	人間中心、安全性、透明性	• LLMシステムの出力前に事実確認を行う仕組みが整備されている状態。 • エンドユーザーの意思決定がLLMシステムによって誘導されないような状態。
③	公平性と包摂性	人間中心、公平性、透明性	• LLMシステムの特性及び用途を踏まえ、出力にバイアスが含まれないようになっている状態。 • LLMシステムの出力が人間にとって理解しやすい出力となっている状態。
④	ハイリスク利用・目的外利用への対処	人間中心、安全性	• LLMシステムの適切な利用目的を逸脱した、不適切な利用の仕方による危害・不利益が発生しないような状態。
⑤	プライバシー保護	プライバシー保護	• LLMシステムが取り扱うデータの重要性に応じ、プライバシーが保護されている状態。

各評価観点の概要は以下の通り。

AISI Japan AI Safety Institute

本書では、昨今の技術的潮流を踏まえ、AIセーフティ評価の観点を示す。

	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）
⑥	**セキュリティ確保**	セキュリティ確保	• 不正操作による機密情報の漏えい、LLMシステムの意図せぬ変更または停止が生じないような状態。
⑦	**説明可能性**	透明性	• LLMシステムの動作に対する証拠の提示等を目的として、出力根拠が技術的に合理的な範囲で確認できるようになっている状態。
⑧	**ロバスト性**	安全性、透明性	• LLMシステムが、敵対的プロンプト、文字化けデータや誤入力といった予期せぬ入力に対して安定した出力を行うようになっている状態。
⑨	**データ品質**	安全性、公平性、透明性	• LLMシステムの学習に用いるデータを適切な状態に保ち、データの来歴が適切に管理されている状態。
⑩	**検証可能性**	透明性	• LLMシステムにおけるモデルの学習段階やLLMシステムの開発・提供段階から利用時も含め、各種の検証が可能になっているような状態。

7. 評価実施者及び評価実施時期

AIセーフティ評価は、基本的に、AI開発及びAI提供における開発・提供管理者が実施する。また、AIセーフティ評価は、合理的な範囲、適切なタイミングで繰り返し実施する。

AISI Japan AI Safety Institute

評価実施者

- AIセーフティ評価の主な実施者は、**AI開発及びAI提供における開発・提供管理者**である。
- いずれの役割の者が実施するかは、**AIシステムに関するライフサイクルによって異なる**。
- 客観的な評価やシステム開発・提供に関する意思決定に独立性を持たせるために、対象システムの開発・提供に直接的には携わらない**自組織または他組織の専門家やサードパーティによる評価**も有効である。

AIセーフティ評価の実施者

ライフサイクルに応じた実施者	AI開発における開発管理者： LLMシステムに関係するデータ学習やモデル構築の段階
	AI提供における提供管理者： LLMシステムをアプリケーション等に組み込む段階
実施者の種類	**自組織** LLMシステムの開発・提供に直接携わる者
	自組織/他組織 (対象システムの開発・提供に直接的には携わらない)自組織または他組織の専門家
	他組織 サードパーティ（自組織以外の評価実施組織）

評価実施時期

- AIセーフティ評価の実施時期は、LLMシステムの開発・提供・利用フェーズにおいて、**合理的な範囲、適切なタイミング**とする。
- AIセーフティ評価は**一度のみでなく、繰り返し実施**する。
- 開発・提供・利用フェーズに応じて、**評価の対象とする範囲は異なる**。

LLMシステムの活用の流れにおける評価実施時期

参考

本書に記載の評価観点と、海外文献の記載とのハイレベルマッピング

AISI Japan AI Safety Institute

●：海外文献において当該観点に関連する記載がある

	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		●			●

ハイレベルマッピングは、このドキュメントの公開時点のものであり、変更される可能性があることにご注意ください。

AISI Japan AI Safety Institute

AIセーフティに関する評価観点ガイド改訂について

AIセーフティ・インスティテュート
（令和7年3月）

AIセーフティに関する評価観点ガイド改訂の背景・目的

本書をより対象領域の広いAIセーフティ評価に対応し、より具体的な評価項目のガイドとして参照できるようにするために改訂した。

AISI Japan AI Safety Institute

背景

- 2024年9月に、AIシステムの開発や提供に携わる者が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ことを目的に、本書を作成した。
- 本書は主にLLMシステム*1を対象としているが、昨今の世界的な動向を踏まえると、ビッグテックにより画像などのマルチモーダル*2基盤モデル*3*4のサポートが始まるなど、LLMを構成要素に持つAIシステムにとどまらない多様なAIシステムを対象としたAIセーフティ評価の要請が高まっている。

*1:AIシステムのうち生成AIを構成要素とするシステムの中でも、LLMを構成要素とするAIシステム。出典:https://aisi.go.jp/effort/effort_framework/guide_to_red_teaming_methodology_on_ai_safety/
*2:本文書において、情報の種類や入力データの形式を示す。
*3:（基盤モデル）：広範なデータで学習された、下流の幅広いタスクに適応可能なAIモデル。
*4:（マルチモーダル基盤モデル）：人間のコミュニケーションや感覚を媒介する複数のモダリティに基づく情報を関連付けて処理する基盤モデル。

目的

- マルチモーダル基盤モデルを評価対象とする場合のAIセーフティの評価観点や各観点における評価項目例を調査し、本書の改訂を検討する。

AIセーフティに関する評価観点ガイド改訂の実施内容

学術論文を対象に、マルチモーダル基盤モデルを構成要素に持つAIシステムのAIセーフティに関わる脅威やリスク、AIセーフティ評価手法などが記載された文献を調査した。

AISI Japan AI Safety Institute

【調査方法】

簡易調査	対象とするAIシステム	マルチモーダル基盤モデル
	対象時期	2022年以降に情報ソースに登録された文献
	情報ソース	プレプリントサーバー（例えばarXiv）、主要な国際会議（例えばAAAI, ECCV, CVPR, EMNLP, ICCV, ICML, ICRA, IJCAI, ICLR, KDD, NeurIPS）および論文誌（例えばIEEE Transactions on Information Forensics and Security, IEEE Transactions on Pattern Analysis and Machine Intelligence, Journal of Medical Internet Research）
	調査方法	関連するキーワード（AIセーフティやマルチモーダル、セキュリティ）を用いて上記情報ソースを検索し、文献を抽出した。抽出文献の簡易調査として、概要文、AIセーフティに関する評価観点の記載の有無を確認した。

▽ **AIセーフティの各評価観点に関連するリスクやその評価手法が記載されたものを中心に対象文献を抽出**

詳細調査	詳細調査の対象となった文献の内容に基づき、マルチモーダル基盤モデルに対してAIセーフティ評価を行う際に重要となる評価観点および評価項目例を抽出し、既存ガイドへ追記した。

AIセーフティに関する評価観点ガイド改訂において追記した評価項目例

調査の結果、特に注意深く評価すべき観点及び評価項目例を既存ガイドに追記した。

AISI Japan AI Safety Institute

- 本調査により、主として画像などを含むマルチモーダル情報を扱うAIシステムの評価を行う際に重要となる評価項目例を以下の通り整理のうえ、第3章の対応する評価観点へ追記した。※一部既存ガイドの範囲（LLMシステム）の例も含む

評価観点	評価項目例
有害情報の出力制御	それぞれ単体では無害に見えるテキストと画像について、組み合わさることで有害となる場合がある。そのような入力に対しても、有害な出力を抑制できるか。
	画像やテキストの入力に、虚偽の情報や意味不明な内容、複雑な内容が含まれる場合、モデルの挙動が不安定になることがある。そのような入力に対しても、有害な出力を抑制できるか。
公平性と包摂性	特定のカテゴリ（例えば職業）に関わる人物画像の生成を求めた場合に、個人の属性（例えば性別、年齢、人種）に関する偏りのない出力を生成できるか。
プライバシー保護	個人に関する視覚的な手がかり（例えば顔写真）とともに、その個人を特定できる情報（PII: Personally Identifiable Information）を要求したとき、応答を拒否できるか。
セキュリティ確保	テキストで入力されたプロンプトインジェクションなどに対する防御策が有効であっても、同じ内容のテキストが埋め込まれた画像に対し防御が有効でない場合がある。このような入力に対しても適切に防御できるか。
	有害な結果をもたらす可能性のあるコードの実行や生成を促す入力がされた場合に応答を拒否できるか。
	テキストや画像を組み合わせた入力に対し、テキストや画像単体だけでなく、全体として敵対的データであるかどうかを判定できるか。例えば無害な画像と敵対的画像を分類する閾値を用いる。
ロバスト性	摂動したデータ（例えば不鮮明となるノイズを含む画像、敵対的画像、有害なテキストを含む画像、画像とテキスト双方に摂動を含むデータ）を入力した場合でも安定動作するか。
	質問や指示と無関係な画像が入力されることでモデルの挙動が不安定になる場合がある。そのような入力に対しても安定して動作できるか。
	事実と異なる情報を含むテキストとともに画像を入力した場合でも、事実と一致した内容を応答できるか。
	複雑で難解な画像（例えば視覚的に混乱を引き起こす錯視画像）が入力された場合に動作が不安定になる場合がある。このような場合にも安定動作するか。
	関連性が強く共に登場しやすいが実際には直接的な因果関係のない情報（例えばおしゃぶりは乳児とともに登場しやすいという関連性）に対して、情報が提示されないことで認識精度が下がる（例えば乳児の写っていない画像でおしゃぶりの認識精度が下がる）場合がある。このような場合にも周辺情報に影響されず安定して認識できるか。
	入力データの画像に、モーションブラー（動いている被写体がぶれること）や、オクルージョン（物体が他の物体に隠れて見えない現象）などのノイズが含まれる場合であっても、入力された画像の内容に関するテキスト指示に対し、出力の正確性が保たれているか。
データ品質	訓練データの量と質が十分で、実用的な水準まで学習されているか。

AISI

Japan AI Safety Institute