AIセーフティ評価環境 検討タスクフォース
検討報告書（2025年度版）
概要

2026年4月15日
AIセーフティ評価環境 検討タスクフォース

AISI Japan
AI Safety Institute

背景・目的

- AIモデルやAIシステムを安全・安心して活用するためにはAIセーフティに基づく評価・対策が不可欠。
- AIセーフティ評価環境の活用方法や機能強化を検討するため、2025年10月にAIセーフティ評価環境検討タスクフォース(以下、検討TF)を設置。
- 本報告書は、2025年度の検討TFの活動内容とその成果を示す。

✕ 野放図に開発・利用される安全でないAI

- ユーザーや社会への悪影響
- AI開発・提供事業者にとってのレピュテーションリスク

✕ 安全ではあるが不便で誰にも望まれないAI

- 実用性の低下によるユーザー離反
- 高コスト化によるビジネス性の喪失

AIセーフティ評価環境によるバランスの良い適切な内容のAIセーフティ評価

AIセーフティ評価環境 検討タスクフォース
活用方法の確立と機能強化を推進

安全・安心で広く利用されるAI

AIセーフティ評価環境 検討TFの体制

- AIセーフティへの関心が高く、AIビジネスに関する知見や開発・運用実績のある国内外の事業者で構成
- 2025年度は評価環境の課題抽出や2026年度以降の機能拡張に向けた検討を実施

TFメンバ

- SB Intuitions株式会社
- NTT株式会社
- 株式会社NTTデータグループ
- NTTドコモビジネス株式会社
- 株式会社Citadel AI
- 日本電気株式会社
- 株式会社野村総合研究所
- 富士通株式会社
- 株式会社Preferred Networks
- Microsoft Corporation
- 株式会社Ridge-i

他

- AIセーフティ評価環境の開発への関心が高い企業で構成
- TFの議論への参加や報告書作成に協力
- リーダー企業は設定せず、フラットに議論

事務局

- AIセーフティ・インスティテュート（AISI）
- TF運営事務局（株式会社三菱総合研究所）

- ファシリテータを担当

AIセーフティに関する評価観点ガイド

AISI Japan AI Safety Institute

- 評価観点ガイドは、AIシステムを開発・提供する事業者が、AIセーフティの観点から自らのシステムが適切な状態にあるかを確認し、維持・向上につなげるための基本的な考え方と評価観点を整理

	評価観点	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	• 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	• LLMシステムの出力に対して事実確認を行う仕組みが整備されている状態。 • エンドユーザーの自律的な意思決定が尊重され、LLMシステムの出力によって安易に誘導されないような状態。
③	公平性と包摂性	• LLMシステムの出力に有害なバイアスが含まれず、個人または集団に対する不当な差別がない状態。 • LLMシステムの出力がすべてのエンドユーザー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	• LLMシステムがハイリスクな目的で利用される場合でも、エンドユーザーやステークホルダーの安全や権利が守られる状態。 • 事前に想定したLLMシステムの利用目的を逸脱した不適切な目的外利用がなされない状態。また、仮に目的外利用された場合にも大きな危害・不利益が発生しない状態。
⑤	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、適切にプライバシーが保護されている状態。
⑥	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	• 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	• 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	• モデルの学習時も含め、LLMシステムがアクセスす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	• モデルの学習段階や、LLMシステムの開発・提供段階・利用時も含め、各種の検証が可能になっている状態。

AIセーフティ評価環境

- AIセーフティ評価環境は、評価観点ガイドやレッドチーミング手法ガイドに基づくAIセーフティ評価を具体化し、AIセーフティ評価の実装を支援することを目的として整備
- 2025年9月に、無償で広く利用可能な形態として、OSSとしてGitHub上に公開

ガイドライン

AI事業者ガイドライン
1.1版(2025/3)，総務省・経産省

AIセーフティに関する評価観点ガイド
1.10版(2025/3)，AISI

AIセーフティに関するレッドチーミング手法ガイド
1.10版(2025/3)，AISI

ツール

OSSとして公開している
AIセーフティ評価環境
(2025年9月初版公開)

AIセーフティ評価環境
・評価ツール
・評価データセット

自動レッドチーミングツール

検討タスクフォースの活動（全体像）

- 各メンバ間でAIセーフティに関する課題認識や評価ツールのフィードバックを共有
- 外部有識者の見解も交えて議論を行い、論点を整理
- 本TFのアウトプットとして、本報告書を作成

TF設置

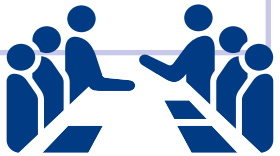
課題認識の共有

AIビジネスにおけるAIセーフティや評価の位置付け、課題を共有

評価ツールの試用・フィードバック共有

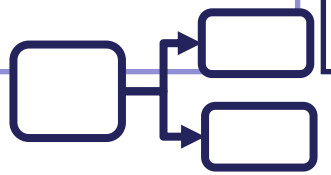
AIセーフティ評価ツールをメンバ各社で試用

改善点や課題を共有



論点整理

AIセーフティ指針や評価ツール・シナリオ活用等に関する論点を整理



アウトプット作成

論点をまとめた検討報告書の作成・公開

2026年度に向けた検討

有識者へのヒアリング

多角的な視点からの具体的示唆

検討タスクフォースで整理した主要な論点

AISI Japan AI Safety Institute

- AIセーフティへのビジネス現場での取り組み課題を解消する**包括的な指針の提示が先に立ち**、その指針を達成する**１つの手段として評価ツールの位置づけを整理**することが重要
- 上位の論点である①と②を優先的に検討し、③は2026年度以降に具体的な要件の検討を行う予定

① **AI普及のための指針**：ビジネス拡大と安全性の両立
各社AIセーフティへの課題意識を踏まえ、ビジネス拡大と安全性両立のために、AIセーフティの取り組みにおける必要条件を整理

② **評価シナリオ/ツール活用**：評価ツールのターゲット再設定と活用
AI普及のための指針と評価ツールへのフィードバックを踏まえ、ツールの開発目的や要求事項を整理

③ **開発要件**：開発目的を達成するために必要な機能要件整理
評価ツールへのフィードバックを踏まえ、機能を整理

検討タスクフォースで整理した主要な論点（AI普及のための指針）

AISI Japan AI Safety Institute

- AIセーフティにおいて、概念や基準の曖昧さ、投資判断のジレンマ、技術進化への対応負担、社会受容性の問題などが企業において**複合的な課題**となっている。
- リスクの整理と基準の明確化、評価手法の共有、アクセラレータとしての設計を通じて、**安全とイノベーションの両立を図ることが求められる。**

概念及び共通枠組みの不透明性	評価基準の曖昧さによる実用性への影響	ガバナンスへの投資判断におけるジレンマ	技術進化への適応と支援の不足	社会受容性への配慮
● AIセーフティ等の概念が統一されず、参照すべき枠組みが多数存在し、安全性評価に関する社内外への必要十分性の説明が困難。	● 評価の到達点が定まっておらず、現場で判断に確信が持てないことにより、生成AI導入の停滞につながる可能性がある。	● AIセーフティへの投資対効果が見えにくいため、AI利活用とリスク管理部門の間で予算や優先順位の調整が難しい。	● 高度化する攻撃手法の情報収集や評価・対策に要する人員やコストの負担が増大。	● 社会がどの程度のリスクを許容するかへの配慮が不可欠であるが、過度な安全を期待すると、有用なイノベーションが失われる可能性
リスク類型や評価の位置づけを整理	**リスクシナリオに応じた最低限の必須基準と推奨基準を明確化**	**AIセーフティを“利活用のアクセラレータ”として設計**	**評価手法・対策に関する具体例の共有** ※公開範囲は配慮が必要	**ルール整備や啓発等を通じて、リスクと便益のバランスを形成**

検討タスクフォースで整理した主要な論点（評価シナリオ／ツール活用）

AISI Japan AI Safety Institute

位置づけ・目的

- **評価環境の目的の明確化とAI活用促進への活用**
  評価環境の位置づけと利用目的を示し、AI活用促進の攻めの側面もあることを意識付けることが重要
- **評価結果による影響の理解と説明責任への活用**
  評価結果の法的な位置づけを意識し、企業が合理的注意を尽くしたことを示す根拠として活用できるように今後進化させることができれば、企業にとっての価値は大きく高まる

評価対象

- **AIシステムの構成に応じた評価**
  評価対象の明確化とアプローチ検討
- **サプライチェーンリスクへの対応**
  モデル更新や設定変更に対応した継続的評価
- **モダリティや新技術への対応**
  マルチモーダル、AIエージェント・フィジカルAIへの対応
- **領域特化の評価アプローチ**
  領域特化型の評価環境を段階的に開発・提供し、実効性を向上

評価内容

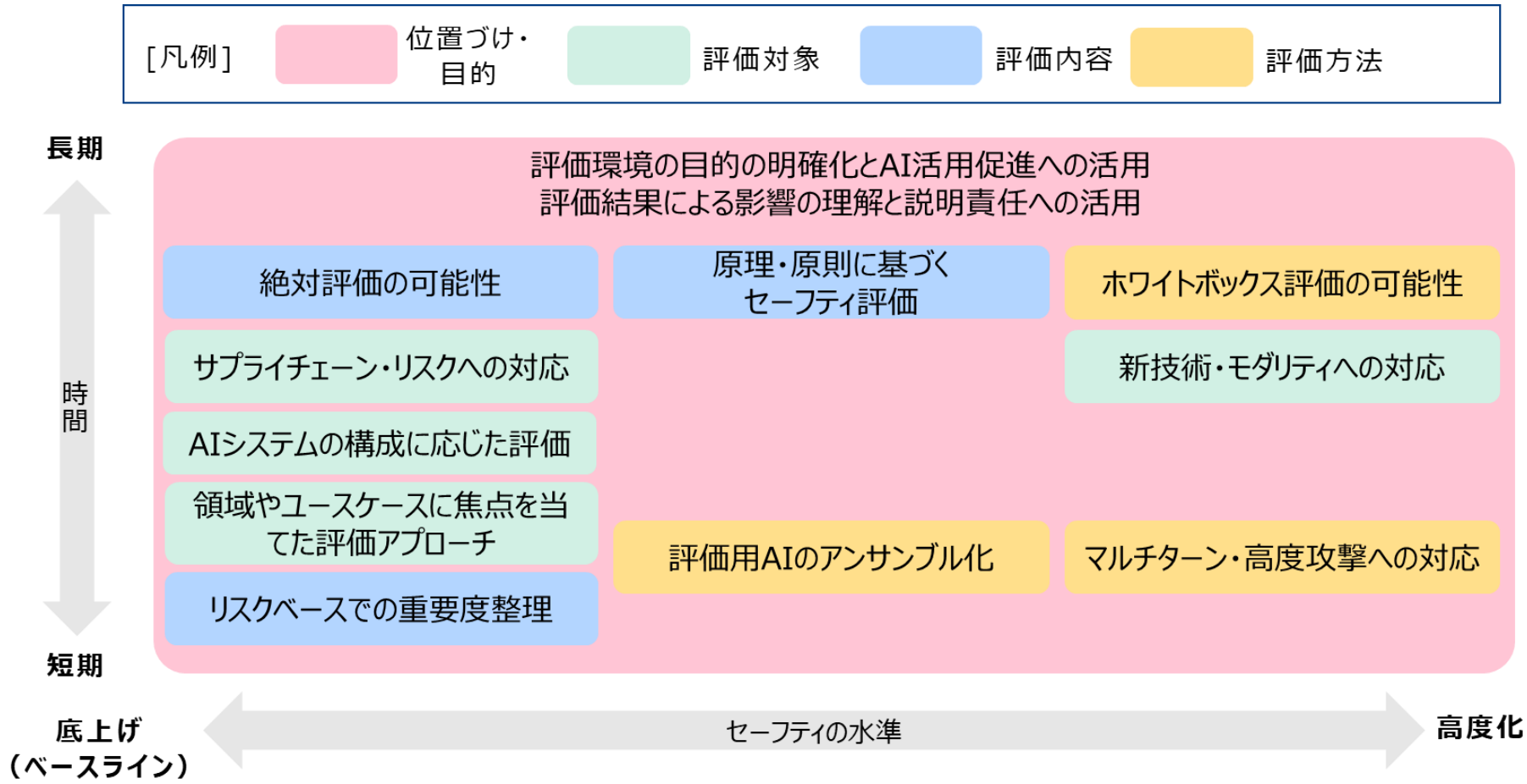
- **リスクベースでの重要度整理**
  最低限守るべき水準と用途やリスクに応じて高度化が求められる水準の整理
- **絶対評価の可能性**
  社内外説明のための一定の基準の設定の可能性（閾値設定の妥当性や社会的合意形成等の課題が伴う。）
- **原理・原則に基づくセーフティ評価**
  個々のプロンプトの表現や攻撃手法に依存せず原理・原則に基づいてセーフティを評価するアプローチ

評価方法

- **マルチターン・高度攻撃への対応**
  Jailbreakや巧妙な誘導による誤動作等への対応策の検討
- **評価用AIのアンサンブル化**
  モデル特有のバイアスやランダム性を回避するための複数の評価用AIを用いた比較・統合等
- **ホワイトボックス評価の可能性**
  モデル内部の情報を活用したホワイトボックス評価の可能性（ただし、技術の完成度や適用容易性に課題あり）

まとめ

- 評価環境の発展に向けた優先度付けやロードマップ検討につなげるため、「時間軸（短期～長期）」と「求められるセーフティ水準（底上げ～高度化）」の2軸で下図の通りマッピング。
- 今後は、多様な論点を基に、技術面やビジネス面での最新動向も踏まえつつ、優先度や機能の具体化の議論を続け、AISIと検討TFと**でAIセーフティ評価環境の機能強化のロードマップを作成していく**計画である。



（参考） 検討タスクフォースで整理した主要な論点（開発要件）

AISI Japan AI Safety Institute

- 設計面やデータセットに関する課題提起があり、2026年度以降に優先度を設定し検討する予定。

設計面

観点	課題
評価フロー設計	段階的に深掘りできる評価設計
結果提示	重要リスク・アラートの優先表示
UI/UX	GUI完結型評価の継続・強化
解釈可能性	定性評価における判定理由の可視化
拡張性	調査用プロンプト等の追加
性能	評価処理速度・実行効率の向上
評価妥当性	複数評価用AIの選択・比較
運用管理	評価結果の管理機能（削除・表示制御）
柔軟性	ベース評価＋オプション評価の切替

データセット

観点	課題
プリセット拡充	用途別プリセット評価データセットの不足
作成支援	データセット自作を支援する補助機能
共有・普及	データセット・ノウハウ共有基盤の構築
持続可能性	データセット汚染回避・更新の運用
ドメイン対応	業界法令に基づくOK/NGデータ
国内特性	日本の公序良俗・年齢別リスク対応
網羅性	評価データの網羅性・多様性確保
自由度	データセット準備・編集の自由度向上

（参考）有識者へのヒアリング

- ビジネス性や法制度等に関する専門家4名へヒアリングを行った。
- 評価環境の位置づけや水準、リスクの階層化などに関する意見があった。

対象者	専門領域・スタンス
佐久間 弘明 氏 (AIガバナンス協会業務執行理事、総務省AIガバナンス検討会 構成員)	【ビジネス性・実用性】 産業界の実装・コンセンサス
羽深 宏樹 弁護士 (京都大学特任教授、スマートガバナンス株式会社CEO・弁護士)	【ビジネス性・実用性、法制度】 アジャイルガバナンス
稲谷 龍彦 教授 (京都大学法学研究科教授)	【法制度】 法理論・社会受容
殿村 桂司 弁護士 (長島・大野・常松法律事務所)	【法制度】 企業法務・紛争実務

得られた主な示唆

- 評価環境の設計において、位置づけと利用目的を明確に示すことと、企業が安心してAIを活用できる環境を整備するという攻めの側面も意識することが重要
- リスクシナリオやドメインごとに目指すべき水準を整理した指針を提供することが期待される。
- 全てのリスクを同等に扱うのではなく、階層化して優先度を明確にすることが求められる。
- 領域特化型の評価ツールを段階的に開発・提供することで、実効性を高められると考えられる。
- 評価結果の位置づけを意識し、企業が合理的注意を尽くしたことを示す根拠として活用できることが望ましい。

（参考）AIセーフティ評価環境検討タスクフォース 総会

AISI Japan AI Safety Institute

- 検討TFのメンバ以外の一般の参加者も募り、前提知識が無くてもAISIや「AIセーフティ評価環境」、検討TFなどについて概要を把握でき、関心を高めてもらうことを目的として開催
- 一般からの参加希望者やオンライン参加も含めた総勢100名が出席

日時： 2026年2月26日（木）13:30～17:10
場所： 株式会社三菱総合研究所5階会議室
（オンライン配信あり）

主な内容：
- AISIの活動紹介
- 「AIセーフティ評価環境」の機能説明
- 検討TFにおける検討状況の報告
- 検討TFメンバをパネリストとしたAIセーフティやAIビジネスに関するテーマについて議論するパネルディスカッション

https://aisi.go.jp/activity/activity_information/260306/

AISI

Japan AI Safety Institute