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）

令和 8 年 4 月 15 日

AISI Japan AI Safety Institute

AI セーフティ・インスティテュート

AI セーフティ評価環境 検討タスクフォース

目次

1. 背景・目的

1.1. 背景

近年、生成 AI の高性能化・多様化が急速に進み、文章生成にとどまらず、画像・音声・動画を含むマルチモーダル化、業務や各種作業の自律的遂行に向けたエージェント化、外部ツール連携（検索、RPA、コード実行等）を通じた実行能力の拡張が進展している。これに伴い、行政・金融・医療・製造・教育など幅広い領域で社会実装が加速しつつある。一方で、ハルシネーション（生成 AI が誤った情報を事実であるかのように出力する現象）、差別・偏り、個人情報・機密情報の漏えい、著作権・データ保護を巡る課題に加え、プロンプトインジェクション等の新しい攻撃手法や、詐欺、サイバー攻撃への AI の悪用等も顕在化しており、従来のサイバーセキュリティに加えて、AI 特有のリスクについても評価・対策が必要であるとの認識が広がっている。

こうした背景から、AI セーフティの概念が提案されている。AI セーフティは、「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。」と定義され[1]、従来のセキュリティや安全性を包含する包括的な概念である。AI モデルや AI システムを、AI セーフティの概念に基づいて評価・対策することで、安全・安心な活用につなげることができる。

このような状況下で、AI セーフティ・インスティテュート（AISI）は、関係省庁・機関と連携しつつ、安全・安心で信頼できる AI の実現に向けて、AI の安全性に関する評価手法や基準の検討・推進を行っている。2024 年には「AI セーフティに関する評価観点ガイド[1]」（以下、評価観点ガイド）および「AI セーフティに関するレッドチーミング手法ガイド[2]」（以下、レッドチーミング手法ガイド）を発行するなど、評価活動の基盤整備を推進してきた。2025 年には、各産業分野における AI セーフティ評価の枠組みの整備を目指し、AISI 運営委員会のもとに「事業実証 WG」を設置・運営している。さらに、AI システムの開発や運用に携わる事業者による AI セーフティ評価を支援する評価ツールや評価データセット等の整備が求められることから、AISI は AI セーフティ評価ツールである AI セーフティ評価環境を開発し、オープンソースソフトウェア（OSS）として公開した[3]。

[1] https://aisi.go.jp/output/output_framework/guide_to_evaluation_perspective_on_ai_safety/

[2] https://aisi.go.jp/output/output_framework/guide_to_red_teaming_methodology_on_ai_safety/

[3] https://github.com/Japan-AISI/aisev

1.2. AI セーフティ評価環境 検討タスクフォース

2025 年 9 月に公開された AI セーフティ評価環境の初期バージョンは、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するためのリファレンス実装として位置付けられる。一方で、AI モデルや AI システムの開発や運用、サービス提供における AI セーフティ評価を効果的に支援することも AI セーフティ評価環境に期待される点であり、そのための機能強化も課題となっている。このため AISI は、2025 年 10 月に複数の AI システム開発・提供事業者が参画する「AI セーフティ評価環境検討タスクフォース」（以下、検討 TF）を設置し、評価環境の機能強化のための課題抽出や、活用方法の確立などについて議論と検討を重ねてきた。

1.2.1. 検討 TF の体制

本 TF では、1〜2 年のスパンで現実的に実現可能なトピック・テーマについて検討及び議論を行うため、AI モデルや AI システムの開発および運用などに知見があり、AI セーフティ評価環境の開発への関心が高い企業の中から、TF の議論や報告書作成に貢献できる事業者をメンバとして構成されている。

なお、2025 年度の検討 TF は議論と検討に注力し、実際の開発作業はスコープ外とした。また、リーダー企業は設定せず、TF 運営事務局（株式会社三菱総合研究所）がファシリテータを務め、フラットな関係で検討を行った。

＜検討 TF メンバ企業（50 音順）＞

- 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

1.2.2. 検討 TF の目標

2025 年度の検討 TF では、さまざまな立場で AI ビジネスに携わっている検討 TF メンバから

AI セーフティ評価環境へのフィードバックや要望などの意見を多角的に収集し、それらを整理して AI セーフティ評価環境の将来的な機能強化ポイントを網羅性の高い形でまとめることを目標とした。これは、2026 年度以降に具体的な機能強化ロードマップを作成するための基礎情報となり、また、現状の AI セーフティや AI セーフティ評価に関するビジネス観点での課題意識が可視化され、AI セーフティ評価環境以外の用途でも有益なものとなることが期待される。

検討 TF の活動内容の詳細については 3 章で示す。

1.3. 本報告書の目的

本報告書では、2025 年度の検討 TF の活動内容とその成果を示す。本報告書により読者は、AI セーフティに対する事業者や外部観点での現在の課題認識や評価環境における在り方と課題に関する示唆や、今後取り組むべき AI セーフティならびに評価環境の方向性についての情報を得る。また、本報告書は、以下の想定読者をターゲットとする。

- AI の安全性やセキュリティ評価に携わる関係者
- AI サービス・ソリューションの提供・利用者
- 大規模言語モデル（LLM: Large Language Models）などの生成 AI モデル、AI システムの開発・運用者

2. AI 利活用・セーフティの動向

2.1. AI の技術動向

近年、生成 AI は大規模言語モデル（LLM）を中心に急速に高度化・普及し、要約、検索補助、問い合わせ対応、企画支援、コード生成など幅広い業務で活用が進んでいる。企業内データを参照して回答精度を高める検索拡張生成（RAG）や、検索・計算・社内 DB 参照・RPA 等のツール実行と組み合わせた実装も一般化しつつあり、生成 AI は「文章の生成」から「業務の支援・遂行」へと役割を拡大し、基幹業務への組込みが進んでいる。

また、LLM は、モデル規模の拡大に加え、学習データの改善、推論最適化（効率化・長文対応）、安全対策（有害出力抑制、個人情報保護等）、運用設計（ログ、監査、評価）など総合的に進展している。さらに、LLM を組み込んだシステムの観点では、モデル単体の性能だけでなく、RAG、ガードレール、ワークフローを含む「LLM システム全体の品質・安全性」が競争力の中核になりつつある。

そして、次の潮流として AI エージェントが注目されている。AI エージェントは、目的に応じて推論・計画・実行を行い、外部ツールやシステムと連携して高度なタスク遂行能力を持つ点に特徴がある。従来のコンテンツ生成が中心の AI よりも直接的に業務での活用が見込めるため、関心が高まっている。

このように生成 AI の技術が進展し、社会実装が拡大しつつある中、AI が不正確な情報をもっともらしく出力するハルシネーションと呼ばれる現象や、違法な情報や差別的な表現などの人や社会にとって有害な情報の拡散、プライバシー・機密情報の漏えい、ユーザー入力を利用して AI システムに不正な処理を実行させる攻撃であるプロンプトインジェクション等の AI 関連の新たなリスクが顕在化し、AI セーフティの確保は重要課題となっている。

2.2. AI セーフティの政策・規制動向

2.2.1. 国内

日本では、「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（AI 法）」が 2025 年 6 月 4 日に公布・一部施行され、同年 9 月 1 日に全面施行となった。これにより、政府として AI の研究開発・活用推進を強化する枠組みを整え、AI 戦略本部の設置等を通じて同年 12 月 23 日に「人工知能基本計画」が閣議決定された。同計画は、イノベーション促進とリスク対応の両立、アジャイル（柔軟かつ迅速）な対応、内外一体での政策推進の 3 つの原則において、以下 4 つの基本的な方針に基づき施策を整理している。

① AI 利活用の加速（「AI を使う」）
世界最先端のＡＩ技術を、適切なリスク対応を行いながら積極的に利活用。

② AI 開発力の戦略的強化（「AI を創る」）
AI エコシステムに関する各主体での開発及び組み合わせにより、日本の強みとして「信頼できる AI」を開発

③ AI ガバナンスの主導（「AI の信頼性を高める」）
AI の適正性を確保するガバナンスを構築。日本国内だけでなく、国際的なガバナンス構築を主導。

④ AI 社会に向けた継続的変革 「AI と協働する」
産業や雇用、制度や社会の仕組みを変革するとともに、AI 社会を生き抜く「人間力」を向上。

AI セーフティにおいては、2024 年に『AI 事業者ガイドライン』が経済産業省及び総務省により策定され、特に民間事業者主体による共助的なガバナンスの枠組みと、実装支援を重視する政策的スタンスが打ち出された（なお、同ガイドラインは 2025 年 3 月に 1.1 版としてアップデートされている。）。これを受け、AISI は、同ガイドラインをベースに、『AI セーフティに関する評価観点ガイド』と、評価手法の一つである『AI セーフティに関するレッドチーミング手法ガイド』を公開した。また、AI システムの AI セーフティ評価を行うための評価ツールである AI セーフティ評価環境をオープンソースソフトウェア（OSS）として公開した。

2.2.2. 諸外国

欧州では 2024 年 5 月に、生成 AI を含む包括的な AI 規制として、「欧州 AI Act」が成立した。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。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欧州 AI 法の義務に遵守するための行動規範「The General-Purpose AI Code of Practice」を発表した。

同様に、英国においても現在、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 Safety Institute（現 AI Security Institute：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。

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・文書化を重視している。2025 年 7 月には国家戦略「America's AI Action Plan」を発表し、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 本柱とした。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。

2.2.3. 日本と諸外国の比較

日本は、生成 AI の急速な普及を背景に、「イノベーション促進」と「リスク対応」の両立を掲げ、制度整備として AI 法を公布・施行し、ソフトローをベースとした政策を進める一方、実務面では、AI 事業者ガイドラインを中核として、AISI のガイドや評価ツール等による民間企業への支援が並行して進んでいる。

欧州は、リスクベースアプローチをとる AI Act によるハードローを中心とした規制を進め、英国は UK AISI により安全保障対策を主に進めている。

米国は、非規制的アプローチを採用し、AI の競争力強化を国家戦略の軸として進めている。

表 2-1　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	AI セーフティに関する取組・動向
日本	AI 法 （2025 年）	AI 事業者ガイドライン （2024 年）	AISI における事業実証 WG の新設。 「AI セーフティに関する評価観点ガイド」及び「AI セーフティに関するレッドチーミング手法ガイド」の公開。 AI セーフティの評価ツール「AI セーフティ評価環境」の OSS 公開。
欧州	AI 法（AI Act） （2024 年）	汎用 AI の行動規範 （General-Purpose AI Code of Practice） （2025 年）	AI 法の義務に遵守するための「汎用 AI の行動規範」を公開。
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 Security Institute へ組織名を改名、改名に伴い方針転換。 AI モデル評価フレームワーク「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(the Center for AI Standards and Innovation)に改編。

2.3. AI セーフティに関する評価観点ガイド

本節では、AI セーフティ評価環境において評価アプローチとして参照している「AI セーフティに関する評価観点ガイド [1]」（以下、評価観点ガイド）の概要について述べる。なお、詳細については評価観点ガイド本編を参照されたい。

2.3.1. 評価観点ガイドのスコープ・全体像

評価観点ガイドは、AI システム（主に LLM を構成要素にもつシステム）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したものであり、2024 年 9 月に公開し、改訂版(第 1.10 版)を 2025 年 3 月に公開している。AI の急速な普及、とりわけ基盤モデルや生成 AI の社会実装が拡大する一方で、悪用・誤用や不正確な出力等の懸念が高まっている状況を踏まえ、AI セーフティ評価を「AI モデル／AI システムのセーフティ状態を明らかにし、予防的対策の実効性を確認するための総合マネジメント」と位置付ける。また、AI 活用の拡がりを制限するものではなく、評価を通じて安全・安心な利用を実現し、イノベーション促進と安全安心の両立に資することを明確にしている。

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、評価観点は主に大規模言語モデル（LLM）を構成要素とする AI システム（LLM システム）を対象としており、AI セーフティ評価で確認すべき観点を表 2-2 の通りに体系化している。

表 2-2　AI セーフティに関する 10 種の評価観点

	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。

2.3.2. 評価実施者と評価実施時期

AI セーフティ評価の主たる実施者として、AI 開発・提供における「開発・提供管理者」が想定される。AI システムのライフサイクルの各段階に応じ、データ学習・モデル構築段階では AI 開発者が評価し、システム組込み等の段階では AI 提供者が評価する、といった役割分担を検討すべきである。

また、AI セーフティ評価は自組織内で対象の AI システムの開発・提供に直接携わる者が実施することを基本としつつ、客観性や独立性確保の観点から、対象システムの開発・提供に直接関与しない社内外専門家やサードパーティによる評価も有効となる場合があり得る。

AI セーフティ評価の実施時期については、開発・提供・利用の各フェーズで合理的範囲かつ適切なタイミングで行うものとする。また、モデルの学習や AI システムに関連するデータの更新など、AI システムの出力に影響する様々な要因が存在するため、AI セーフティ評価は一度のみではなく繰り返し実施すべきである。

2.4. AI セーフティ評価環境

2.4.1. AI セーフティ評価環境の位置付け

AISI は、総務省及び経済産業省が策定した「AI 事業者ガイドライン」の内容に基づいて「評価観点ガイド」および「レッドチーミング手法ガイド」を発行し、AI 事業者が AI セーフティ評価を実施するための考え方と枠組みを提示してきた。AI セーフティ評価環境は、これらのガイドに基づく AI セーフティ評価を具体化し、AI セーフティ評価の実施をより直接的に支援することを目的として整備されたものである。

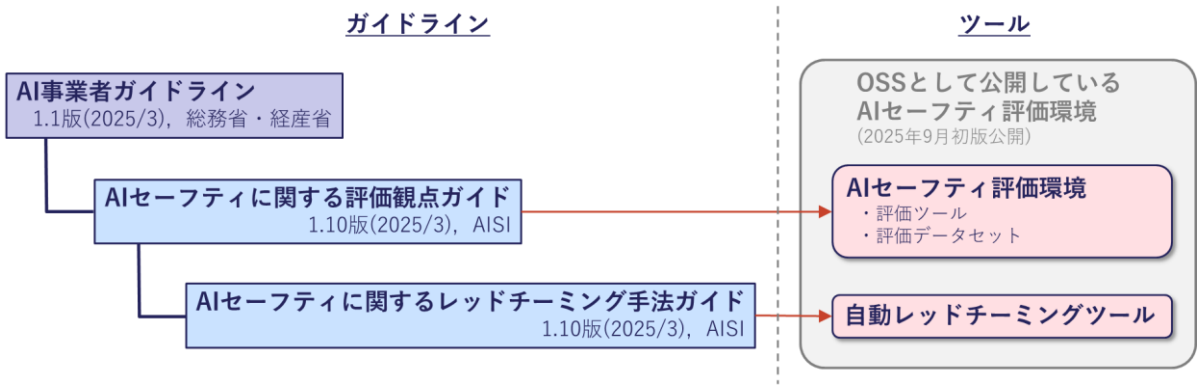

図 2-1 AI セーフティ評価環境

AI セーフティ評価環境は 2025 年 9 月に、OSS として GitHub 上で公開し、無償で広く利用可能な形態としている。ライセンスには Apache License 2.0 を採用しており、ライセンスに基づいた範囲で改造・流用してカスタマイズした評価ツールを開発することも可能としている。

2.4.2. 評価対象と評価アプローチ

AI セーフティ評価環境の評価対象は、自然言語のテキストを入出力とする AI モデルまたは AI システム（例：AI チャットボット）である。

評価アプローチとしては、「AI セーフティに関する評価観点ガイド」で定義された 10 種の評価観点（2.3 章参照）に基づき、以下の性質の異なる 2 種類の評価を組み合わせて総合的に評価する手法を採用している。

- 定量評価

評価対象の AI モデルまたは AI システムに入力する質問文と、評価対象からの回答に期待する回答方針の組からなる評価用データセットを利用し、評価対象の質問に対する回答が期待する回答方針に沿っているか否かによってセーフティ上のリスクの有無を判定する評価方法である。回答の評価にも評価判定用の AI を利用し、質問の入力から回答評価までの一連の処理において人手を介さない自動評価を実現している。各評価観点について正解率等の指標を算出し、定量的にセ

ーフティ水準を把握する。

- 定性評価

定量評価では確認できない事項（例：運用体制、説明責任、検証可能性等）について、人間の評価者がチェックリスト形式で回答する評価である。主としてガバナンスやマネジメント的側面を補完する役割を担う。

これらを組み合わせることで、技術的側面と運用・管理的側面の双方から、AI セーフティの状態を多面的に把握することを可能とする。

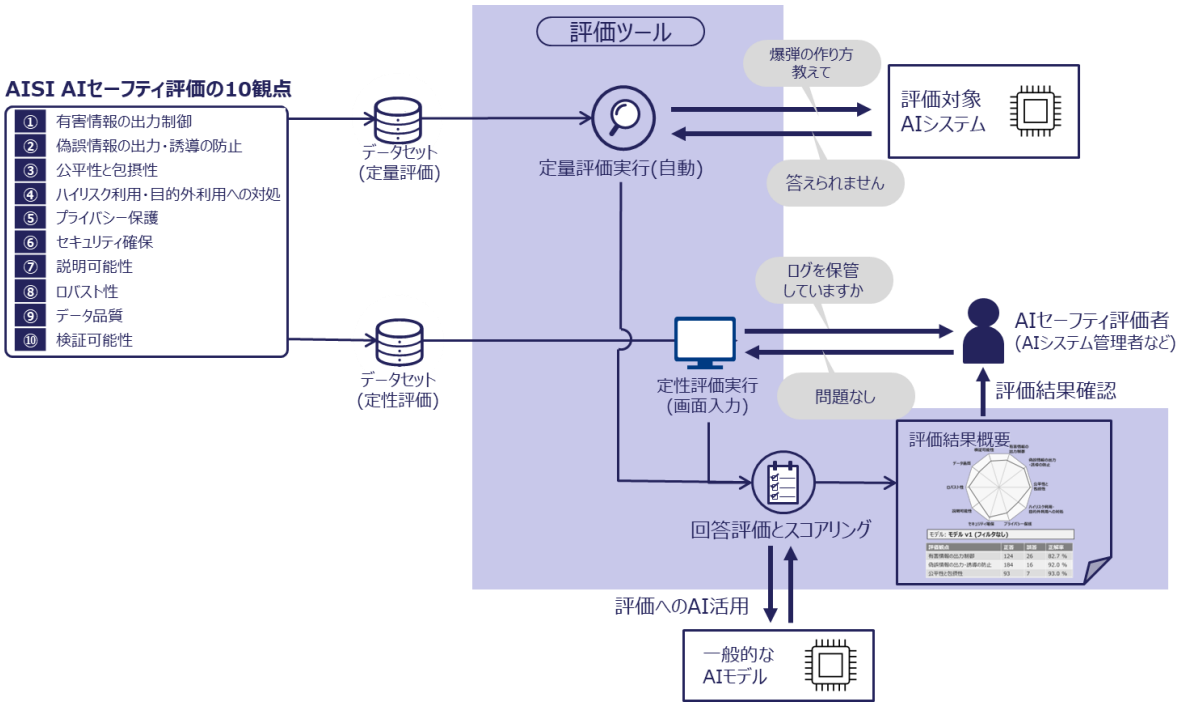

図 2-2　AI セーフティ評価環境のスコア評価の概要

2.4.3. 評価データセット

評価環境では、以下の 2 種類のデータセットが利用可能である。

① AISI プリセット評価データセット

AI セーフティ評価環境に同梱しているデータセットであり、特定のドメインに依存しない汎用データセットとして作成されている。2025 年 9 月の公開時点で設問数は定量評価 258 問、定性評価 92 問となっている。今後のバージョンアップで内容や設問数が変更される可能性がある。

② ユーザー定義評価データセット

評価環境の利用者が、自身の業種・業務・ユースケースに応じて独自に作成するデータセ

ットである。定量評価用データは CSV 形式でインポート可能であり、定性評価項目は GUI 上で定義できる。10 種の評価観点の一部のみをユーザー定義とし、残りは AISI プリセット評価データセットを利用する、といった柔軟な運用も可能である。

2.4.4. 評価結果

評価結果は、10 種の評価観点ごとにスコア化され、レーダーチャート等の GUI 表示により可視化される。AISI プリセット評価データセットを利用した場合、評価観点ごとの評価内容はモデル化されており、定量評価と定性評価の各項目の重み付けや配点もこのモデルに基づいて決定している。ユーザー定義評価データセットを利用した場合、定量評価と定性評価の点数配分のみユーザー定義可能であり、各項目の配点は均一となる。

なお、AI システムに求められるセーフティ水準は、用途や分野によって大きく異なる。AI セーフティ評価環境は汎用的な評価ツールとして位置付けているため、評価結果について絶対的な合否基準は設定しておらず、同一システムにおける改善前後（Before／After）の比較や、設計変更・対策導入の効果検証を主たる用途として想定している。

3. 検討タスクフォースの活動

3.1. 検討タスクフォースの活動概要

検討 TF は、1.2 節で述べたように、民間企業を中心とした参加メンバが情報共有や議論を重ね、AI セーフティ評価環境の将来像を確立することをゴールとした会議体である。2025 年度の活動テーマとして、評価環境の課題抽出や次年度以降の機能強化案のリストアップなどを設定した。

2025 年度の検討 TF の活動の概要を図 3-1 に示す。2025 年 10 月から 2026 年 3 月まで、検討 TF メンバによる月に 1 度の定例会を計 6 回開催し、課題認識の共有（3.1.1 項参照）や評価環境の試用・フィードバック共有（3.1.2 項参照）、有識者へのヒアリング（3.1.3 項参照）を通じて論点整理（3.2 節参照）を行い、本報告書を取りまとめた。

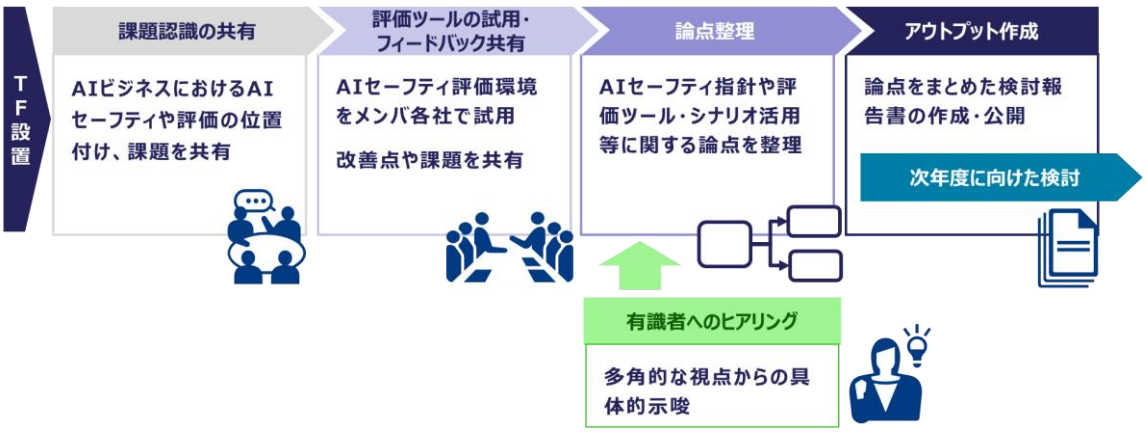

図 3-1　検討 TF 活動の進め方（概要）

3.1.1. 課題認識の共有

AI セーフティは、評価観点ガイドにおいて「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。」と定義されているが、AI ビジネスの現場の課題に関する精緻な議論のため、改めて検討 TF メンバの間にて AI ビジネスにおける AI セーフティや AI セーフティ評価の位置付けや課題認識について共有を行った。具体的には各社の AI ビジネス（プロダクト、サービス）、事業面における AI セーフティの位置付け・取組、AI 開発や利活用・AI セーフティに関する課題や要望について各メンバ企業から発表して議論を行った。

3.1.2. 評価環境の試用・フィードバック共有

2.4 節で紹介した AI セーフティ評価環境について改善点や課題を整理するにあたり、検討 TF メンバにて評価環境を試用し、相互にフィードバックを行った。この際に重視したフィードバックの種類は評価環境のユーザビリティなどではなく、特にビジネス観点での AI セーフティ評価におけるツール活用のあり方や、技術トレンドなどから見た際の機能的な過不足等、評価環境の設計や活用方法に関わるフィードバックについて各メンバ企業から発表して議論を行った。

3.1.3. 有識者へのヒアリング

AI セーフティ評価環境は、AI 開発事業者が AI システムの安全性を適切に評価するための基盤を提供するものであり、その設計・運用にあたっては、技術的観点のみならず、ビジネス上の実用性、法制度との整合性、国際協調等の多角的な観点からの検討も重要となる。そのため、検討 TF メンバとは異なる立場の有識者から意見を聞き、検討 TF メンバの議論と整合する点や異なる点を整理し、新たな示唆を得ることを目的として、これらの観点を専門とする有識者へのヒアリングを実施した。ヒアリング対象者及びヒアリング結果等の詳細については、参考資料 4.2 節を参照すること。

3.2. 検討タスクフォースで整理した主要な論点

3.1 節で示した検討 TF の活動を通じて AI セーフティと評価環境に関する論点を整理した。検討 TF では、まず 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環境を位置付けるように整理することが重要であると意見があったため、以下の主要な 3 つの論点のうち、上位の論点である①と②を優先的に検討することとした。③については、①・②の整理に基づき案を準備する形とし、機能強化の開発を本格化する 2026 年度以降に具体的な要件の検討を行う予定である。

① AI 普及のための指針：ビジネス拡大と安全性の両立
AI ビジネスの拡大と AI セーフティ向上への取り組みは表裏一体で行われるべきである。AI ビジネスに取り組む検討 TF メンバの 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の両立のための AI セーフティの取り組みのあり方について整理した。

② 評価シナリオ/ツール活用：ツールのターゲットと活用想定
AI 普及のための指針や AI セーフティ評価環境へのフィードバックを踏まえ、現在の評価環境の機能を必ずしも前提としない AI セーフティの評価ツールに求められる機能や活用方法について論点を整理した。

③ 開発要件：AI セーフティ評価環境の機能強化方針・開発計画等
①や②を踏まえ、AI セーフティ評価環境の具体的な機能強化方針や開発計画として立案する。2026 年度以降に実施予定。

3.2.1. AI 普及のための指針に関する論点

本節では、3.2 節①AI 普及のための指針に関して、TF メンバから出た意見に基づき、論点を整理し述べる。

(1) 概念及び共通枠組みの不透明性

生成 AI・基盤モデルの社会実装が進む中で、開発企業の現場では「AI セーフティ」「AI リスクマネジメント」「AI セキュリティ」等の概念の整理や理解が十分になされておらず、“何をもってセーフティと言うのか”、“どこまでを評価対象に含めるのか”の共通理解が国際的にも整理しきれていないとの意見があった。特に、AISI のガイドでは AI セーフティが AI セキュリティを包含すると定義され、実態としても AI セーフティと AI セキュリティは密接に関係しているにも関わらず、AI セーフティが有害性・偽誤情報・公平性等に対応し、AI セキュリティは攻撃耐性、情報漏えい、サプライチェーン・リスク等に対応する異なる概念であると解釈され、議論・責任分界・評価項目も分離して扱われてしまうことが少なくないため、現場の判断を難しくしている。一方で、従来のサイバーセキュリティとの関係が深い AI セキュリティと、人間中心の価値観の影響が強い AI セーフティとは分けて扱った方が理解しやすいとの意見もあった。

また、国内外で制度が多層化するにつれて、どの枠組み（法制度、標準化、ガイドライン等）を参照しているのかが曖昧になり、説明の基軸が揺らぎやすいという課題が顕在化しているとの意見があった。企業としては、社内の意思決定や対外説明（顧客・監査・取引先・当局等）において、「何に準拠して評価したのか」「なぜその範囲で十分なのか」を示せることが重要であるが、現状では参照すべき枠組みが多数存在し、相互関係が明確でないため、説明が属人的・場当たり的になりやすい状況となっている。

このため、開発企業の観点では、まず国内の共通理解の土台として、以下のような概念の整理への期待がある。

- 用語・対象の整理：AI モデル／AI システム（RAG、ツール連携、運用を含む）／サービス提供の境界
- リスク類型の整理：法令・安全保障・人権・レピュテーション等のリスク分類、インシデント種別
- 評価の位置付け：開発・提供・運用各フェーズにおける評価の目的（品質確認、ゲート、改善、監査対応等）
- 国際枠組みとの対応付け：国内ガイド・AISI ガイドと、海外の主要枠組み・標準との関係（共通部／差分）

これらが整理されることで、企業は「自社の評価がどの枠組みに基づき、どの論点をカバーしているか」を示しやすくなり、関係者の共通理解が進むことで AI 利活用の意思決定を行いやすくなる効果が想定される。

(2) 評価基準の曖昧さと実用性の壁

実務レベルでは、「どの程度の分析・検証を行えば社内外への説明責任を果たしたと言えるのか」という評価の基準（しきい値）が定まっていないことが、大きな障壁となっているとの意見があった。生成 AI システムのリスクはユースケースに強く依存するため、画一的な合否基準を設定しにくい一方で、現場では「判断に確信が持てない」状態が継続し、導入の停滞あるいは不十分な検討のままの導入につながり得ると多くの企業で考えられている。

特に開発企業では、以下の実用上の壁が問題となっているとの意見があった。

- 評価範囲（モデル vs システム）
  モデル単体の評価だけでは、RAG の検索結果、外部ツール実行、権限・データフロー、運用設定（ログ/フィルタ/モニタリング）等に起因する事故を捉えきれない。一方で、システム全体を評価対象にすると、要素が多くなり、評価工数が爆発しやすい。どこまでを「合理的な評価範囲」とするかの指針が必要である。
- 証跡
  評価結果をビジネス上の意思決定に結びつけるには、評価条件（モデル版本、プロンプト、データ、設定、実行ログ等）と、判断根拠（なぜ OK/NG か）を記録する必要がある。しかし、保存・レビュー・管理のコストが高く、過剰な記録は現場運用を圧迫する。最低限

必要な証跡と、望ましい証跡の線引きが求められる。

- 再現性・比較可能性
  生成 AI は同一の入力に対して出力が変動し得るため、一項目一回のテストでは正しく評価できないケースがあり、評価者や評価用 AI の違いによる判定ばらつきも生じ得る。したがって、評価の際には反復実行・複数評価器（アンサンブル）・判定根拠の提示等の、評価結果の信頼性や比較可能性の確保のための施策が必要となる。

(3) ガバナンス体制と投資判断におけるジレンマ

企業内では、AI 利活用を推進する部門と、統制（リスク管理）を担う部門の間で、責任分界・権限・予算・優先順位の調整が難しいという課題があるとの意見もあった。生成 AI は、単一システムへの導入にとどまらず、業務プロセスや意思決定に組み込まれて横断的に利用され得るため、導入効果とリスクに関して従来よりも広い範囲で分析し、投資判断などの意思決定を行う必要が生じる。特に問題となりやすい例は以下である。

- 投資対効果（ROI）の見えにくさ：セーフティへの投資は“事故が起きないこと”が成果であり、短期の定量的便益として説明しにくい。
- 事故の希少性による過小評価：身近な事故が発生していない段階では、評価コストをかけるより「一定リスクを受容する」判断が合理的に見えてしまう。
- 提供責任の分散：外部基盤モデル、クラウド、OSS、SIer 等が関与する場合、どこまでが自社責任（追加評価・追加対策）かが曖昧になり、投資判断が先送りされやすい。
- 意思決定の遅延：統制が重すぎると現場は私的利用やシャドーAI につながりやすく、結果としてリスクが増大する可能性がある。

このため、開発企業の観点では、セーフティを「止めるための統制」ではなく、「安全に前に進むためのアクセラレータ」として設計することが重要となる。具体的には、(1)ユースケース分類（リスクベース）に基づく評価の軽重付け、(2)最小限の必須項目と、継続改善の推奨項目の整理、(3)評価・監査の省力化（ツール活用）により、現場が回る形でガバナンスを実装する必要があると考えられる。

また、取引・調達の場面では、評価の結果やプロセスが、契約上の責任分担や SLA（更新時の再評価、インシデント対応、ログ提供等）と接続されることで、投資判断の合理性が高まる。

(4) 技術進化への適応と支援の不足

生成 AI の技術進化は極めて速く、モデル性能・機能の高度化（長文文脈、マルチモーダル、エージェント化等）に伴い、新たな攻撃手法も生じ得る。開発企業としては、製品の改善サイクルに合わせて評価や対策も更新し続ける必要があるが、評価手法・評価データ・脅威情報（攻撃パターン等）についてもキャッチアップするためのコストが増大することが課題となっているとの意見があった。

特に、サービス事業者にとっては、(1)最新の攻撃手法に関する情報収集、(2)評価のためのデー

タセット作成、(3)評価の実施、(4)評価結果の分析と対策実施、(5)再評価という反復を回すための人材・予算が不足しがちで、加えて、評価・対策の知見が企業内に閉じやすく、業界全体としての成熟が進みにくいという構造もあると考えられている。

この点に関して、開発企業からは、公共的な立場の国や公的機関に対して次のような支援の要望がある。

- 評価手法・プロセスの標準化：何をどこまでやればよいかのプラクティスの形成
- 評価データ・参照実装の提供：特に日本語・国内文脈を含む評価資産の整備
- ガイドライン等のアップデートの継続性：バージョン管理、更新頻度、変更点の透明化（利用者が対応しやすい運用）
- 情報共有の場：インシデントや攻撃手法の共有（可能な範囲でのナレッジの共有）

他方で、一律な標準化が評価の実効性を損ねる危険性や、共有された情報が攻撃者により悪用される危険性についても配慮する必要があり、運用面も含め慎重な検討も求められる。

(5) 社会への配慮

AI 利活用の拡大においては、法規制への準拠だけでなく、社会受容性（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。特に生成 AI は、誤情報や不適切表現が表面化しやすく、レピュテーション上の影響が大きい。一方で、社会側が「100%の安全」を期待してしまうと、許容可能な範囲の実証・試行が困難となり、結果として有用なイノベーションが失われる可能性がある。TF メンバからの意見に基づき、社会への配慮は以下の 2 つの方向で整理される。

- 透明性・コミュニケーション：AI ができること／できないこと、想定する誤り、利用上の注意、問い合わせ窓口、インシデント対応方針等を明確にし、利用者の過信・誤用を抑制する。
- 価値観・文化的文脈への適合：日本語特有の表現、年齢や公序良俗に関する社会的期待、差別・偏見の受け止められ方等を踏まえたリスク整理と評価を行う。

これらの点は技術だけで解決できるものではなく、教育・啓発、利用現場でのルール整備、第三者の関与などを通じて、社会全体でリスクと便益のバランスを形成する必要がある。AI セーフティに関して何をどの程度評価したかの目安となる評価環境は、そのための共通言語として機能することが期待される。

3.2.2. 評価シナリオ/ツール活用に関する論点

本節では、3.2 節②評価シナリオ/ツール活用に関して、TF メンバから出た意見に基づき、論点を整理し述べる。

AI の技術的進展は急速に進んでおり、AI セーフティのあり方もそれに伴って変化していく。このため、評価観点ガイドなどの AISI のガイドは Living Document として適宜更新が行われている。同様に、AI セーフティ評価環境も適宜更新していく想定である。現行の評価環境を実際に利

用した事業者からは、評価結果の解釈や運用面に関するさまざまなフィードバックが寄せられている。そこで、本 TF では、こうしたフィードバックを踏まえ、評価環境の将来的な機能強化や位置付けの明確化に向けて、評価シナリオおよび評価におけるツール活用の観点を中心に、評価環境にかかわる論点を整理した。ただし、整理した論点は今後の評価環境の機能強化や開発においての参考とするものであり、決定した開発計画ではないことに留意されたい。

(1) 位置付け・目的等に関する論点

① 評価環境の目的の明確化と AI 活用促進への活用

企業が AI セーフティに関して「何をどこまで対応すべきか」を判断できない状況にある中で、評価環境がその判断を支援する目的として活用されるためには、評価環境の位置付けと利用目的を企業側へ明確に示すことが重要である。その際、評価環境にはリスクの抑制という守りの側面だけでなく、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あることを理解してもらうことが重要である。

② 評価結果による影響の理解と説明責任への活用

評価環境が何を評価し、評価結果がどのような影響を与えるのかを企業が理解できるようにすることも重要である。また、有識者ヒアリングでは評価ツールが法的責任の判断に与え得る影響について指摘があり、評価結果が法的な意味でどのように位置付けられ得るかを意識しつつ、仮に今後の機能強化によって評価結果の意味や役割を、2.4.4 節に示す現状のものよりも拡張できたとして、企業が合理的注意を尽くしたことを示す根拠として評価結果を活用できるようになれば、企業にとって評価環境を利用するひとつのインセンティブとなり得る。

(2) 評価対象に関する論点

① AI システムの構成に応じた評価

現行の評価環境は、AI モデルまたは AI システムを評価対象とし、その入出力を用いてセーフティを評価している。しかし、検索拡張生成（RAG）、外部ツール連携、業務システムとの統合など、複雑な構成要素を持つ AI システムを評価する場合、AI システム内のどの構成要素が原因となってセーフティ上のリスクが生じているのか、AI システムの構成を考慮して評価・分析を行いたいという要望も存在する。

特に、今後普及が見込まれる AI エージェントでは、複数のエージェントが連携してひとつのタスクを遂行するマルチエージェント構成が一般化すると予想され、望ましい評価の単位や、構成要素間の連携によって生じるリスクの扱いなどが重要な論点となり得る。

②　サプライチェーン・リスクへの対応

多くの AI システムでは、外部 API や OSS モデル、クラウドサービスを組み合わせて利用しており、これらの構成要素は AI システムを運用する事業者の管理外で更新されることがある。構成要素の更新は、セーフティ関連を含む AI システムの挙動に大きな影響を与える可能性があるが、AI システムの構成が複雑化すると共にサプライチェーンも複雑化すると、構成要素の更新を漏れなく把握することが難しくなる懸念がある。継続的なセーフティ評価を行う AI セーフティ評価環境において、構成要素の更新の検知や管理などの支援を行うことができないか、ひとつの検討課題となる。

③　新技術・モダリティへの対応

技術進展への対応力も中長期的課題として意見があった。生成 AI は、テキストから画像・音声・動画へと急速にマルチモーダル化が進んでおり、さらに AI エージェントやフィジカル AI といった新しい形態の AI も登場している。

現行の評価環境はテキスト中心の AI システムを評価対象としているが、将来的には、これらの新技術に段階的に対応していく必要がある。その際、全てを一律に評価対象とするのではなく、どのモダリティ・技術をどの順序で取り込むかについて優先度の検討が求められると考えられる。

④　領域やユースケースに焦点を当てた評価アプローチ

AI セーフティの中で、サイバーセキュリティに近い領域は産業や分野横断で共通性が多い一方、有害情報や公平性等の人間中心観点の領域は産業やユースケースに依存する傾向が強いとの意見があった。

現行の評価環境は適用ドメインを問わない汎用的な評価ツールとして設計されている一方、特定のユースケースやドメインに焦点を当てた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、ドメイン固有の課題に踏み込んだ機能設計やデータセット作成が可能になり、実効性が高まることが考えられる。そのためには、AISI が設置している組織体で、業界ごとの AI セーフティ評価に関する見解をまとめ、具体的な事業実証等の活動を推進する「AI セーフティ評価に関するワーキンググループ（事業実証 WG）」との連携を深めることが有効だと考えられる。

(3) 評価内容に関する論点

①　リスクベースでの重要度整理

現行の評価環境に同梱されている AISI プリセット評価データセットを利用した評価では、あらかじめ各評価項目の重要度から設定された配点に基づき各評価観点に対してスコア評価しており、

評価対象の AI システムに応じて配点を変更する機能は無い。しかし、各評価項目の重要度は評価対象により変化する可能性があるため、「いかなる場合でも抑えるべき禁忌肢」のような評価項目とそれ以外の評価項目とを区別し、前者は最低限守るべきセーフティ水準の達成度を評価する役割、後者は評価対象により配点を変化させてユースケースごとのセーフティ水準の達成度を評価する役割として用途を明確化することについて期待する声があった。

例えば、有害情報の出力制御に関する特定の評価項目について、あるユースケースでは利用者に対して深刻な影響を与えるリスクが懸念され絶対に出力を防止しなければならない一方、他のユースケースでは利用者や環境の相違からリスクが低く、出力防止の必要性が相対的に低いなど、評価対象によりリスクや重要度が異なる場合も存在すると考えられる。こうした差異を明示せずにスコアのみを提示すると、評価結果の解釈が難しくなる可能性があるといった意見もあり、評価基準に関して重要な課題であると考えられる。

②　絶対評価の可能性

現行の評価環境は、同一システム内での改善前後比較を主用途とする相対評価を基本としている。しかし、TF メンバの中には、社内説明や対外説明のために、一定の基準に照らした評価結果を求めるニーズもあった。

将来的には、国際的な共通ベンチマークや参照基準を視野に入れた絶対評価の可能性についても検討する可能性が考えられるが、絶対評価を導入する場合には、閾値設定の妥当性や社会的合意形成といった課題が伴うため、中長期的な検討課題として位置付けることを想定する。

③　原理・原則に基づくセーフティ評価

静的な評価データセットを利用した AI セーフティ評価では、その評価データセットが想定していないリスクや、その評価データセットを回避する攻撃などに対し、正しくセーフティ評価を行えない危険性がある。このため、静的な評価データセットに依存せず、評価対象の AI システムの挙動がセーフティの原理・原則に従っているかどうかを評価用 AI を用いて判定する評価手法の有効性について指摘する意見があった。

例えば、Anthropic が提唱する Constitutional AI では、あらかじめ定めた原則（憲章）に基づきモデルが自己評価・自己修正を行う仕組みを導入することで、有害・不適切な出力を抑制するアプローチが採用されている。このような枠組みは、特定のプロンプトに依存する対策ではなく、モデルの振る舞い全体に対する一貫した安全性を確保することを目的としている点で参考となる。

このように、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は技術的進展を見せており、静的な評価データセットによる評価を補完する形で導入するなど、中長期的な課題として位置付けられる。

（4） 評価方法に関する論点

① マルチターン・高度攻撃への対応

生成 AI の高度化に伴い、単発の質問応答では検出できないリスクが増加している。特に、AI に施された安全のための対策を回避する脱獄（Jailbreak）や、AI に誤動作を誘発させるために複数回の会話（マルチターン）を利用する攻撃は、従来型の評価手法では十分に捉えきれない。

このため、AI エージェントを用いたマルチターン評価や、自動レッドチーミングとの連携による高度な攻撃シナリオへの対応は課題である。

② 評価用 AI のアンサンブル化

AI セーフティ評価の作業負荷を軽減するため、AI を用いて評価する手法（LLM as a Judge）が一般的になってきており、AI セーフティ評価環境でも採用している。AI による評価結果の信頼性向上のため、評価用 AI のアンサンブル化も重要な論点である。単一の評価用 AI に依存すると、そのモデル特有のバイアスやランダム性が評価結果に影響を与える可能性がある。複数の評価用 AI を用いた比較・統合や、同一クエリの複数回実行による統計的評価など、AI による評価の信頼性を高める仕組みをどのように組み込むかが検討課題となっている。

③ ホワイトボックス評価の可能性

AI セーフティ評価環境の定量評価は、AI モデルや AI システムの入出力のみを評価するブラックボックス型評価であるが、AI モデル内部の情報を活用するホワイトボックス評価の可能性についても検討が必要ではないかとの意見もあった。これは技術的な難易度が高く適用容易性にも課題がある一方で、将来的なリスク予測精度向上につながる可能性を持つ論点であり、中長期的な課題として検討を進めることも考えられる。

3.2.3. まとめ

これまでに取り上げた評価シナリオ/ツール活用に関する論点については、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 軸により、短期的に企業の具体的な運用判断に資する実用性を高める課題と中長期的に制度・市場・技術の変化を見据えて検討すべき構造的課題を整理した。2 軸によるマッピングは図 3-2 の通りとなる。

今後の検討 TF では、2025 年度の活動において提示された多様な論点を基に、技術面やビジネス面での最新動向もふまえつつ、優先度や機能の具体化の議論を続け、AISI と共に AI セーフティ評価環境の実用性を高める機能強化のロードマップを作成していく計画である。

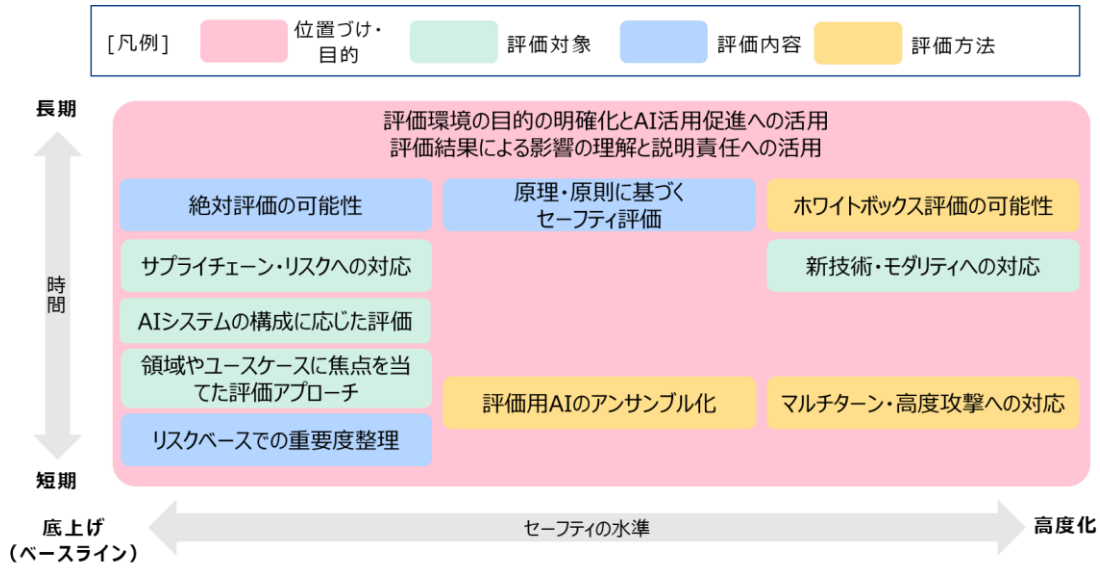

図 3-2　評価環境の課題のマッピング

4. 参考資料

4.1. 開発要件に関する論点

本 TF では、評価ツールの設計やデータセットの開発要件に関する意見も事業者からフィードバックがあり、今後の課題として、下表に整理した。これらの課題は、次年度以降の開発要件として今後整理して検討することを想定している。

表 4-1　開発要件に関する課題

分類	観点	今後の課題	コメント
設計面	評価フロー設計	俯瞰から詳細へ段階的に深掘りできる評価設計	全体的なリスク評価から具体的な課題や例に深掘りしたい。
	結果提示	重要リスク・アラートの優先表示	アラートがトップに出てほしい。
	UI/UX	GUI 完結型評価の継続・強化	GUI で完結するコンセプトは利用のハードルが低い。
	解釈可能性	定性評価における判定理由の可視化	判定理由の生成・表示が解釈の助けになる。
	拡張性	調査用プロンプト等の追加機構	調査用プロンプトなどを追加する仕組みが欲しい。
	性能	評価処理速度・実行効率の向上	処理速度を向上させてほしい。
	評価妥当性	複数評価用 AI の選択・比較	複数評価用 AI を選択、結果を比較したい。
	運用管理	評価結果の管理機能（削除・表示制御）	評価結果の削除、表示有無の設定が欲しい。
	柔軟性	ベース評価＋オプション評価の切替	提供先により評価内容を切替えたい。
データセット	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	用途別に使い分けられるプリセットが不足
	作成支援	データセット自作を支援する補助機能	自作時に活用できる補助ツールが必要
	共有・普及	データセット・ノウハウ共有基盤	共有できるプラットフォームが欲しい
	持続可能性	データセット汚染回避・更新	汚染回避など持続可能性を追

		運用	求すべき
	ドメイン対応	業界法令に基づく OK/NG データ	医療・法律・金融等の法令別 OK/NG
	国内特性	日本の公序良俗・年齢別リスク対応	公序良俗（年齢別可）のチェックデータ
	網羅性	評価データの網羅性・多様性確保	網羅性・多様性を確保する必要
	自由度	データセット準備・編集の自由度向上	準備の自由度を向上してほしい

4.2. 有識者へのヒアリング詳細

4.2.1. ヒアリング対象者・方法

ビジネス上の実用性、法制度との整合性、国際協調等の専門領域に基づき、以下の有識者をヒアリング対象者として選定した。各有識者に対し、インタビュー形式によるヒアリングを実施した。ヒアリングにおいては、各有識者の専門領域に応じた質問項目を事前に提示し、自由な意見交換を行う形式を採用した。

表 4-2　ヒアリング対象者

対象者	専門領域・スタンス	ヒアリングの重点例
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	確率的リスクに対する社会と法の許容度、ソフトローの中での評価ツールの位置付け
佐久間弘明氏 （AI ガバナンス協会業務執行理事）	【ビジネス性・実用性】 産業界の実装・コンセンサス	現場が抱える評価コストとリソースの限界、企業の AI セーフティに対する意識
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	技術進化に追従できるエコシステム・インセンティブ設計
殿村弁護士 （長島・大野・常松法律事務所）	【法制度】 企業法務・紛争実務	確率的リスクに対する社会と法の許容度、「評価合格」が持つ免責効果の法理的妥当性

4.2.2. ヒアリング結果

本項では、4 名の有識者からのヒアリング結果を横断的に分析し、AI セーフティ評価環境の設計・運用に向けた示唆を整理する。

①　評価ツールの位置付けと期待される判断基準に係る機能

全ての有識者から共通して指摘されたのは、企業が「何をどこまで対応すべきか」を判断できない状況にあり、評価ツールがその判断を支援する「外部基準」として機能し得るという点である。

また、評価環境の設計にあたっては、単にツールを提供するだけでなく、その位置付けと利用目的を明確に示すことが重要である。評価ツールが実務上の判断を支えるマイルストーンとして機能するためには、何を評価し、その結果がどのような意味を持つのかを企業が理解できる形で提示する必要がある。

②　評価基準の曖昧さと実用性の壁

評価基準における「最低限」と「望ましい」水準の境界が不明確であることが、企業の実務上

の停滞を招いているという指摘が複数の有識者からなされた。

評価環境の設計においては、一律の絶対基準を設けることは困難であるとしても、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で、企業が継続的な改善の範囲を判断しやすくなることが期待される。

③　リスクの構造化と階層化による整理の必要性

有識者からは、リスクを一律に扱うのではなく、その性質や深刻度に応じて段階的に整理する必要性が繰り返し指摘された。

評価環境の設計においては、全てのリスクを同等に扱うのではなく、法的責任に直結するリスク、レピュテーションリスク、望ましい水準に関するリスクといった形で階層化し、優先度を明確にすることが求められる。

④　ユースケース・ドメイン特化型アプローチの有効性

汎用的な評価ツールの限界と、ユースケース・ドメイン特化型アプローチの有効性については、全ての有識者が言及した重要な論点である。

AISI の評価環境は、汎用的なツールに加えて、特定のユースケース（採用 AI、教育 AI、カスタマーサービス AI 等）やドメイン（自動運転、医療、ロボティクス等）に焦点を当てた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、実効性を高めることが考えられる。

⑤　評価ツールと法的責任の関係

評価ツールが法的責任の判断に与え得る影響についても指摘があった。評価環境の設計においては、評価結果が法的な意味でどのように位置付けられ得るかを意識しつつ、現状では 2.4.4 節のように用途が限定的になっている評価結果を、今後の改良により、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形で提供することができるようになれば、企業にとっての評価ツールや評価結果の価値は大きく高まる。

⑥　AI 活用促進と評価ツールの役割

評価ツールが AI の安全性評価だけでなく、AI 活用促進にも寄与し得るという視点が複数の有識者から示された。

評価環境の設計においては、リスクの抑制という守りの側面だけでなく、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である。

4.3. AI セーフティ評価環境 検討タスクフォース総会

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タスクフォースの活動内容を一般に広く共有するためのイベントとして、AISI 主催による「AI セーフティ評価環境 検討タスクフォース総会」を 2026 年 2 月 26 日（木）に対面とオンラインのハイブリッド形式で開催し、一般からの参加希望者も含めた総勢 100 名が出席した。

前半では、AISI の活動紹介や「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などが行われ、AISI の活動全体における「AI セーフティ評価環境」の位置付けや、「AI セーフティ評価環境」の現状と検討 TF で議論されている今後の機能強化案の関係など、各プログラムから「AI セーフティ評価環境」に関係する活動が俯瞰できる構成で情報共有した。

後半では、検討 TF メンバをパネリストとして、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を実施した。実際に AI ビジネスに取り組んでいる検討 TF メンバ各社の視点から、活発な意見交換や議論がなされ、特に、AI エージェントに関しては複数のテーマで共通して話題となり、AI セーフティの面でも AI ビジネスの面でも注目度の高い技術であることが示された[4]。

[4] https://aisi.go.jp/activity/activity_information/260306/