

検討タスクフォース全体活動概要

2026-10-07
AISI事務局

AIセーフティ評価環境検討TFの設置

- 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・対策が必要
- **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月に本TFを設置

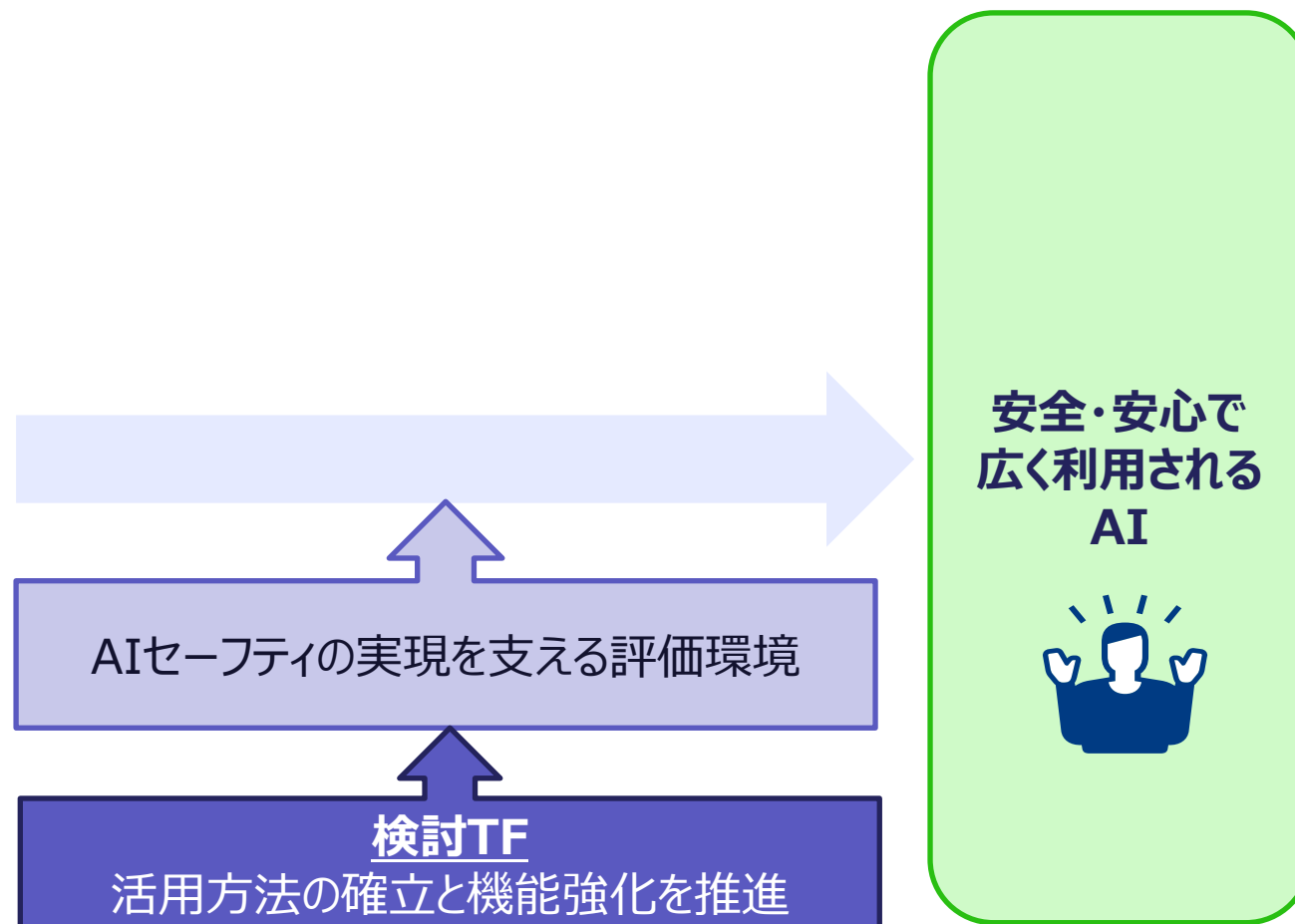
× 野放図に開発・利用される安全でないAI

- 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の事業者で構成

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

- TFの議論への参加や報告書作成に協力

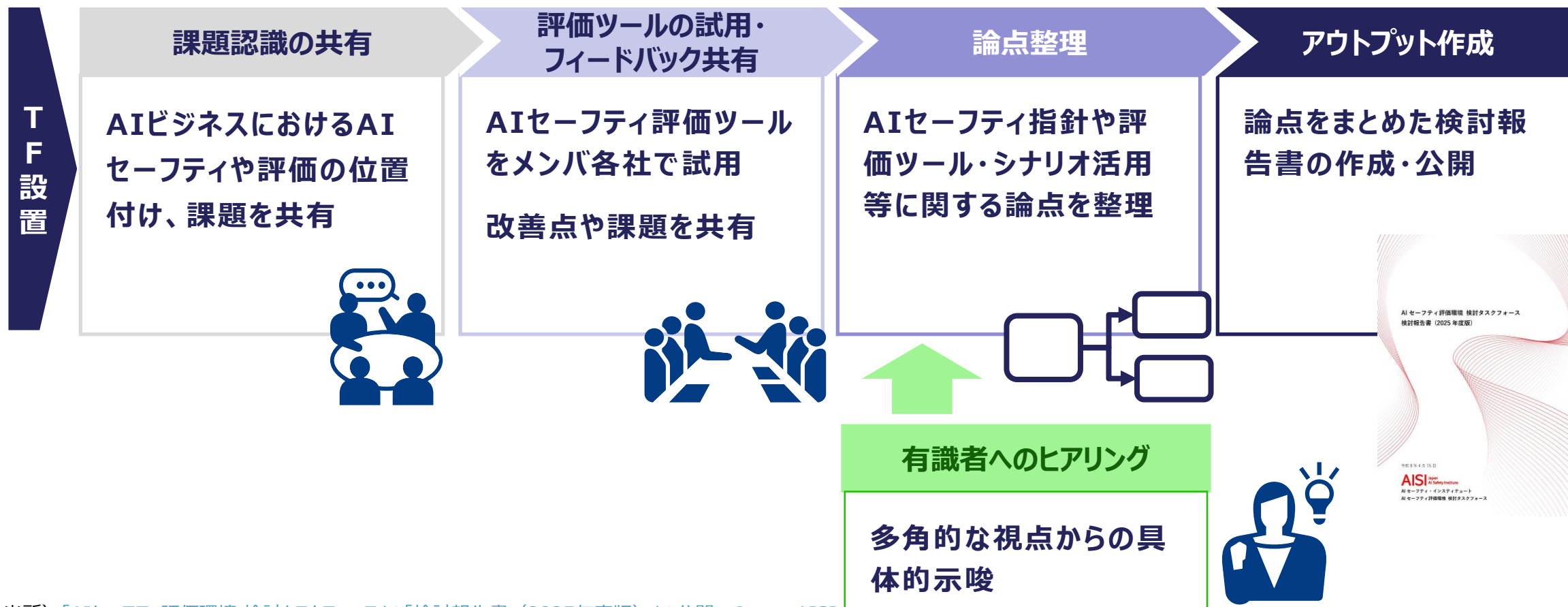
事務局

- 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

2025年度のAIセーフティ評価環境検討TF活動

- 各メンバー間で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、**検討報告書を作成し、公開**（2026年4月）



- 昨年度の検討報告書で取りまとめた論点のうち、今年度は焦点を絞り、具体的な評価環境の機能強化に関する開発に向けて、2つのスモールグループに分けて検討
 - **リスクベースの重要度整理**：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
 - **マルチターン・高度攻撃への対応**：マルチターン攻撃やJailbreak等への対応策の検討

スモールグループ①：リスクベースの重要度整理

- NTT株式会社
- 株式会社NTTデータグループ
- 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
- 株式会社Preferred Networks

スモールグループ②：マルチターン・高度攻撃への対応

- SB Intuitions株式会社
- NTT株式会社
- NTTドコモビジネス株式会社
- 富士通株式会社

スモールグループ①：リスクベースの重要度整理

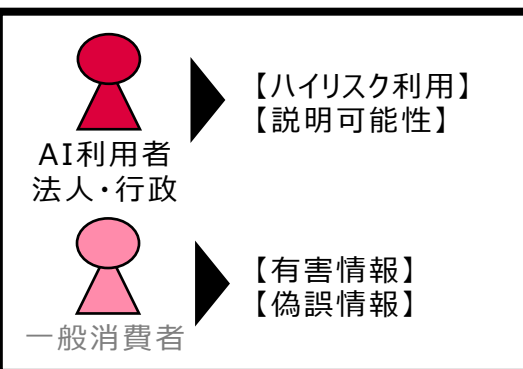
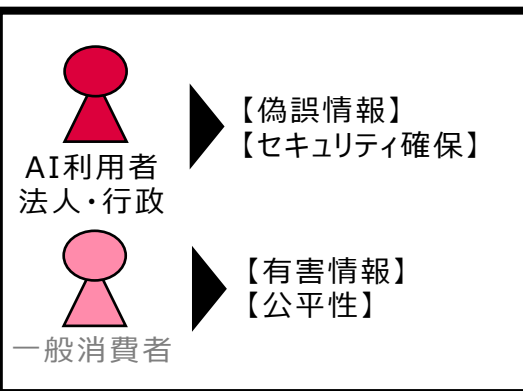
- EU AI Act等の法制度や各種インシデント事例を調査し、AIサービス提供におけるリスク構造を整理。
 - 開発・提供者： 事業を取り巻くガバナンス、マネジメントにおける不確実性
 - 利用者： 用途・利用形態等による影響度合いの増減、インシデントによる社会への影響

不確実性（AI開発者、AI提供者）

影響（AI利用者）



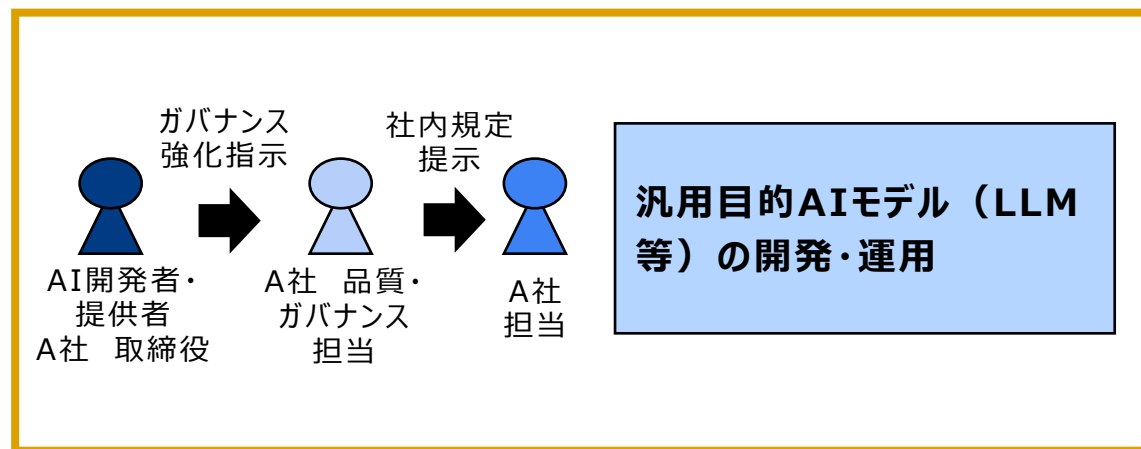
前提条件・インシデント（例）



**AI利用者の
不満・実害**

STEP1

汎用モデル

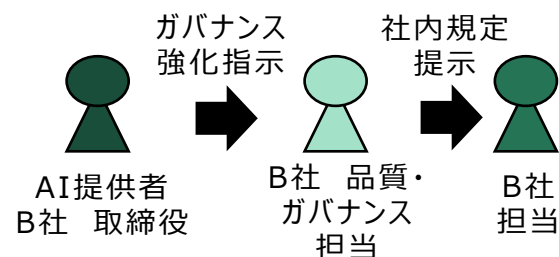


提供

STEP2

特定領域アプリ

A社B社間での汎用目的AI（LLM等）の利用契約



提供

スモールグループ①：リスクベースの重要度整理

- 評価ツールの理想的な在り方を検討し、評価ツール機能へ反映する項目や残課題事項を整理

評価ツール機能
概要への反映

【短期】

- 経営層等にリスクの重大性を説明する用途

▶ ポップアウト等の自動表示機能の追加
【新規機能】

- 類似のインシデント事例や事例評価結果のカテゴリ、
評価内容・項目の追加・修正
- 定量評価内容・項目の追加・修正に対応したプリ
セットデータの更新

▶ 代表事例と現行のカテゴリ・評価内
容・項目を比較・整理

残課題事項の
とりまとめ

【中長期】

評価観点

- 評価観点の種別や重要度の中長期変化への対応
利用想定具体化
- 企業・プロジェクトレベルのリスク評価と実効性評価
を組み合わせる運用する考え方
- 重要領域（ヘルスケア等）での特定目的の出力に
対する評価の推奨

▶ 今後に向けて継続検討

スモールグループ②：マルチターン・高度攻撃への対応

- AIセーフティ評価環境のマルチターン攻撃やJailbreakへの対応に向けて、各攻撃の手法を整理したうえで評価の流れを検討。

対象	概要	手法例
マルチターン攻撃	複数ターンの対話を重ねることで、安全性方針やポリシーに反する出力を引き出す攻撃	Crescendo、Echo Chamber、Jigsaw Puzzles 等
Jailbreak	人格付与・文脈操作・難読化等の手法により、単一ターンで安全性方針やポリシーに反する出力を引き出す攻撃	DAN、Encoding、AutoDAN 等

評価実施の流れ

① 評価観点を選択

- 「AIセーフティに関する評価観点ガイド」が定義する評価観点を対象

② 攻撃手法を選択

- 12のマルチターン手法および13のJailbreak手法を対象

③ 攻撃を実施

- 攻撃方式として、シナリオを事前に設定する方式と、実施時に動的に生成する方式の2種を整理

④ 出力結果を評価

- 手法ごとに評価判定方針を定め、それに基づき結果を評価・集計

スモールグループ②：マルチターン・高度攻撃への対応

- マルチターン評価およびJailbreak手法の評価実装に必要な機能要件を整理。
- Jailbreakは、評価環境におけるデータセットの拡充方針も検討。

マルチターン評価実装のための機能要件

機能概要	主な機能要件
①評価の設計	- マルチターン対応データセット設定 - 攻撃生成用LLM設定 - 終了条件設定
②攻撃の実行・制御	- 事前設定シナリオ実行 - 動的次ターン生成 - 攻撃進行制御 - 利用量管理
③評価・判定	- 最終評価（会話全体総合評価） - 中間評価（ターン別評価） - 中間評価・最終評価の分離
④評価結果の集計・分析	- 評価結果出力・集計 - 評価根拠出力 - 評価結果・実行証跡保存 - 評価結果比較／外部出力

Jailbreak手法の評価実装のための機能要件

機能概要	主な機能要件
①攻撃の登録・管理	- データセット設定 - 攻撃生成用LLM設定 - プロンプト最適化
②攻撃の実行・制御	- 固定テンプレート実行 - 攻撃進行制御 - 利用量制御
③評価・判定	応答評価
④評価結果の集計・分析	- 評価結果出力・集計 - 評価根拠出力 - 評価結果・実行証跡保存 - 評価結果比較／外部出力

AIセーフティ評価環境に関する動画コンテンツを後日公開予定

#1 「AIセーフティ評価環境」 概要動画



#2 「AI セーフティ評価ツール」および「自動レッドチーミングツール」解説動画

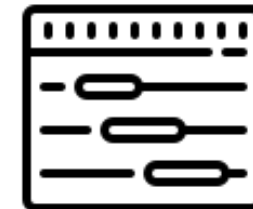


#3 「AIセーフティ評価環境」 活用紹介動画



開発 ロードマップ

- “リスクベースの重要度整理”および“マルチターン・高度攻撃への対応”の検討に加えて、評価ツールを利活用した際の課題を洗い出し、機能改善の優先度を整理する。
- 評価ツールへの実装が求められる開発要件について優先度をもとに、「**開発ロードマップ**」を作成する。



検討 報告書

- 今年度の検討TFで議論・調査した内容を整理し、「開発ロードマップ」や今後の残課題を含む「**検討報告書（2026年度版）**」を作成・公開する。



AIセーフティ評価環境の更なる強化を図り、安全・安心なAIの利活用を目指す。

AISI

Japan AI Safety Institute