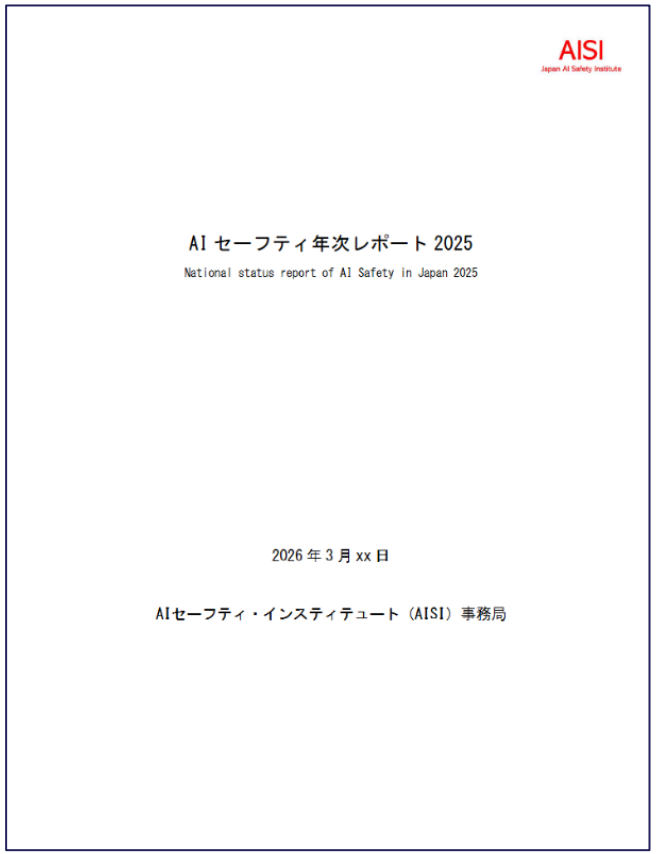
AISIの活動状況（2025年度）

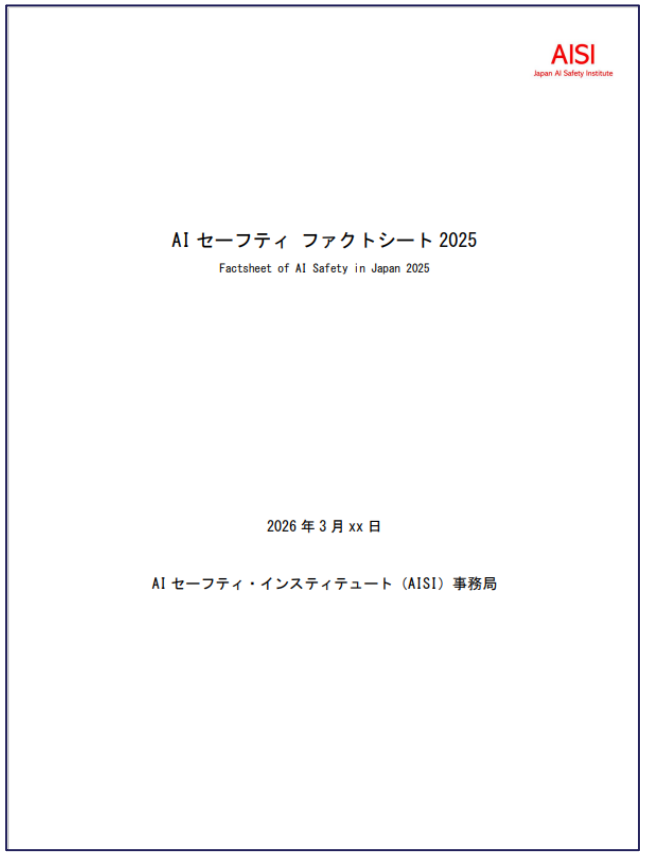
資料１-２

AISI Japan AI Safety Institute

AISIの活動状況を「AIセーフティ年次レポート」及び「ファクトシート」としてとりまとめ。2026年4月に公開予定。

- AISIの2025年度の活動を「AIセーフティ年次レポート2025」としてとりまとめるとともに、AI セーフティに関する公開された文書や組織・体制を中心に国内外の事実関係を記載した「AIセーフティファクトシート2025」を参考資料として公開予定。





活動内容①

AISI Japan AI Safety Institute

AIセーフティ・インスティテュート（AISI）の昨年度の取り組み

＜評価手法・ガイドの体系化＞

● AIセーフティに関する評価観点ガイド

マルチモーダルに関する調査結果や評価項目例などを追記。

● AIセーフティに関するレッドチーミング手法ガイド

試作AIシステムでの実施結果を踏まえて、詳細手順やレポート等を追加。

● AIセーフティの評価ツール（AI セーフティ評価環境）の開発と OSS 公開

上記のガイドに基づくAIセーフティ評価を行うための評価ツールとデータセットをオープンソース・ソフトウェア（OSS）として一般公開。

● 評価環境検討タスクフォースによる官民連携での評価ツールの開発検討

AIセーフティの評価環境に関するフィードバック収集や今後の方向性について議論・検討を行う作フォースを組成。

活動内容②

AISI Japan AI Safety Institute

AIセーフティ・インスティテュート（AISI）の昨年度の取り組み

● AIインシデントレスポンス・アプローチブック

AIシステム特有のリスクに起因するインシデントに対応するための新たな枠組みとして「AI-IRS（AI Incident Response System）」を示したアプローチブックを公開。

● AIシステムに対する既知の攻撃と影響

AIシステムに対する特有のセキュリティ攻撃を俯瞰し、AI及びAIシステムに対する攻撃とその影響をとりまとめ、公表。

● AIシステムのためのデータ品質マネジメントガイドブック

データとAIの価値を最大化するために必要なデータ品質を確保するためにすべきことをガイドとしてとりまとめ、公表。

活動内容③

AISI Japan AI Safety Institute

AIセーフティ・インスティテュート（AISI）の昨年度の取り組み

＜国際連携の深化＞

● 「Hiroshima Global Forum for Trustworthy AI」の開催

AIの安全性、セキュリティ、信頼性等の向上に向けた国内外の最新動向や、各国の安全性／セキュリティ担当機関による取り組みについて議論。

● AISI国際ネットワーク

2025年２月パリ会合（第２回）以降、共同テスト演習を通じて各国と評価手法・知見の共有を進め、７月バンクーバー会合（第３回）で評価レポートを公表。2026年２月インド会合（第５回）の合意文書として「AI評価に関するコンセンサスと課題」を公表。

● アンソロピック社とのMoC

2025年10月、米アンソロピック社との間で信頼できるAIエコシステム構築に向けた協力に関するMoCに署名。

活動内容④

AISI Japan AI Safety Institute

AIセーフティ・インスティテュート（AISI）の昨年度の取り組み

＜産業・社会への展開＞

● Chief AI Officerガイドの公開

組織内のAIガバナンスの枠組み構築に向け、CAIOを設置・運用する際の標準的な実務指針を提供することを目的としてガイドを公表。

● 事業実証ワーキンググループの設置

AIセーフティ評価の活動を広く一般に普及させ、AIの利活用を促進することを目的として設置し、４つのSWG（ヘルスケア、ロボティックス、データ品質、適合性評価）を組成し、分野別ガイド等を作成。

● ベンチマークプロジェクト

AIの安全性評価を具体化し、ベンチマークとして構築・提案することを目的として、利用場面、内容、対策等の分類毎に検討を実施。