AISI
Japan AI Safety Institute

AI セーフティ年次レポート 2025

National status report of AI Safety in Japan 2025

2026 年 3 月 31 日

AIセーフティ・インスティテュート（AISI）事務局

目次

1　AIセーフティ・インスティテュート（AISI）の設置およびこれまでの取り組み

1．1　基本情報

AISI は、安全、安心で信頼できる AI の実現に向けて、AI セーフティに関する評価手法や基準の検討・推進を行うための機関である。AISI は、「AI セーフティ」を通じて、AI の利活用におけるリスクへの対応とイノベーション促進を両立させていく。

AI セーフティ

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

1） AISI の目的（役割）

- 政府への支援として、AI セーフティに関する調査、評価手法の検討や基準の作成等の支援を行う。
- 日本における AI セーフティのハブとして、産学における関連取組の最新情報を集約し、関係企業・団体間の連携を促進する。
- さらに、他国の AI セーフティ関係機関と連携することで、国際的なコンセンサスを確立する。

2） AISI の特殊な状況

AI は、これまでの技術分野では考えられないほど技術変化が速い最先端の分野であり、その影響範囲は技術だけにとどまらず、社会や雇用、文化など広範にわたる。

各国 AISI の置かれている状況はそれぞれ異なり、AI セーフティの対象範囲も、AI システム自体のセキュリティから、エコシステム全体のガバナンス、法律などの整備から、知的財産や個人情報への配慮など多岐にわたっている。米国では、さらに安全保障などへも対象が広がっている。

月や四半期で製品が置き換わるほど技術開発スピードが速く、分野横

断かつ、グローバルな取り組みであり、各国や国際機関のレポートも毎月のように公表されている。こうした中で、日本を含め世界中で一つ一つの取り組みに、これまで以上に時間がかかり、まさに、アジャイル思考で走りながら考えている状況である。従来の年単位で変化する安定した技術に対する対応ではない、全く新しい技術政策の推進の在り方が求められている。数が多いだけではなく、中の質となるような技術の繋がりのチェックも含めて、対応が求められる。

AI セーフティに関する国内横断的やグローバル連携の取り組みは、これまで前例となるものがなく、各国でも政策と技術を融合した体制の作り方を模索している。また、各国が相応の資金を投じる中で、各国の体制の作り方は異なり、その体制の違いが調整をより一層難しくしている反面、AISI は政府機関の一部として活動を行うことで、国として AI セーフティに関する政策を直接推進できる。

各国の政策と技術の推進体制

米国：国務省と商務省と米国国立標準技術研究所（NIST）がチームで取り組みを推進

英国：府省内に AISI を作り、政策と技術の意思決定を一本化

欧州委員会：法律を所管するとともに、利活用も一体で機構内のチームで推進

日本：内閣府と関係省庁が連携して AI の安全性確保に向けた政府方針等を AISI に指示し、AISI はそれに基づいて安全性評価に係る調査、基準等の検討や、実施手法の検討等を行う。

１．２　主要なイベントとアクション

国際的・技術的環境の変化を踏まえ、AISI は 2025 年度において、①評価手法・ガイドの体系化、②制度・標準・認証への対応、③国際連携の深化、④産業、社会への展開への対応、を戦略的重点として活動を展開した。

1）　評価手法・ガイドの体系化

AISI は、AI セーフティを抽象的理念にとどめることなく、実務に適用可能な評価手法やガイドを体系的に整備することを重視している。

① AI セーフティに関する評価観点ガイド(第 1.10 版)の公開（改訂・マルチモーダル観点の拡充）

AI セーフティを評価する際の観点（AI セーフティに関する評価観点）を整理したガイドを、技術動向の変化を踏まえて改訂し、2025 年 3 月に第 1.10 版として公開した。改訂では、マルチモーダルに関する調査結果や評価項目例の追記などを行った。評価設計や社内外説明に使える共通言語・チェック観点として提供することで、事業者が安全性評価を実務に落とし込みやすくする狙いがある。

https://aisi.go.jp/output/output_framework/guide_to_evaluation_perspective_on_ai_safety/

② AI セーフティに関するレッドチーミング手法ガイド(第 1.10 版)の公開（実施例・別紙/別添を加えた改訂）

2024 年に公表したレッドチーミング手法ガイドを、事業者での実践を後押しするため改訂し、2025 年 3 月に第 1.10 版として公開した。改訂では試作 AI システムでの実施結果を踏まえ、詳細手順やレポート等を別紙として追加した。単なる概説に留めず「どうやって回すか」を具体化することで、現場で再現可能な運用手順として活用できる形にしている。

https://aisi.go.jp/output/output_framework/guide_to_red_teaming_methodology_on_ai_safety/

③ 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

AI システムの開発や提供に関わる事業者が、「AI セーフティに関する評価観点ガイド」や「AI セーフティに関するレッドチーミング手法ガイド」に基づいた AI セーフティ評価を実践するには、ガイドで説明された内容を具体的な評価項目やデータセットに落とし込んだ上で AI システムの評価テストを行う必要があり、そのための専門知識や試行錯誤の過程が必要となる点が AI セーフティ評価活動の普及促進のための課題であった。このため、ガイドに基づく AI セーフティ評価を具体化した評価ツールとデータセットのリファレンス実装として、AI セーフティ評価環境を開発し、ツールのソースコードやデータセットを含めてオープンソース・ソフトウェア（OSS）として一般公開した（2025 年 9 月）。

AI セーフティ評価環境は、無償の評価ツール・データセットとして利用可能な他、ソースコードやデータセットを参考にすることで、様々な事業者が自身の環境や条件に合わせてカスタマイズした評価ツ

ールやデータセットを開発することを支援する用途も想定している。

更に、各国の AISI や政府機関などとの議論や技術交流において、我が国の AI セーフティに関する立場や考え方を、ガイドと共にツールも提示することで、より具体的に主張することも可能となり、今後の国際連携においても役割を担うことが期待される。

https://aisi.go.jp/output/output_information/250912/

https://github.com/Japan-AISI/aisev

④ AI セーフティ評価環境 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

2025 年 9 月の AI セーフティ評価環境の公開後、約 10 社の AI ビジネスに取り組む民間企業が参加し、AI セーフティ評価環境のフィードバック収集や今後の機能強化の方向性などについて議論・検討する「AI セーフティ評価環境 検討タスクフォース」（以下、検討 TF）を組成した（2025 年 10 月）。

検討 TF では、各参加企業が自身の AI ビジネスにおける課題意識と現状の AI セーフティ評価環境の整合点やギャップなど、多様な意見交換と議論が行われ、AI セーフティや AI セーフティ評価環境の課題や今後あり得る機能強化の方向性など、2026 年度以降に具体的な機能強化案を確立するための基になる情報が 2025 年度の成果物となった。

また、2025 年度の検討 TF の議論や成果を、検討 TF 参加企業以外にも広く共有するため、一般参加可能な公開イベントである「AI セーフティ評価環境 検討タスクフォース総会」を開催した（2026 年 2 月）。オンライン・オフライン含めて約 100 名が参加し、AISI の活動や AI セーフティ評価環境の機能、検討 TF の検討状況など様々な情報を AISI や検討 TF 参加企業から発信することで、評価環境や検討 TF の認知度・関心度の向上を図った。

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

⑤ AI インシデントレスポンス・アプローチブックを公開

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した。本書では、AI インシデントは起こり得るものとの前提に立ち、被害を最小化する発見的統制の重要性を明確にしている。AI-IRS は、AI サプライチェーン全体を視野に入れたコンセプトを示しており、AI を信頼できる社会基盤として活用するた

めのアプローチを提供している。（2026 年 1 月公開）
https://aisi.go.jp/activity/activity_security/260109/

⑥ 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すべく、学術論文等で発表された AI や AI システムに対する攻撃とその影響を「AI システムに対する既知の攻撃と影響」としてまとめ公表した。（2025 年 3 月）
https://aisi.go.jp/output/output_security/known_attacks_and_impacts/

この資料は、既に公開した「AI セーフティに関するレッドチーミング手法ガイド」に記載の AI システムへの代表的な攻撃手法を補完するレッドチーミングへの活用に加えて、AI セキュリティに関する調査検討や研究開発にも参照できる。総務省「AI セキュリティ分科会」（第 1 回）においても参考資料として採用された。
また、より詳細な内容について、学術論文「Securing AI Systems: A Guide to Known Attacks and Impacts (arXiv:2506.23296)」として公表した。（2025 年 6 月）
https://arxiv.org/abs/2506.23296

⑦ 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の公開
AI は、便利で効率的なサービスを提供し、社会の変化を加速させる。AI の活用を促進するには、AI を安心して活用できるための信頼の確保が不可欠といえる。AI にとってデータはその基盤であり、正しいデータを学習や処理することで正しいデータを導き出せる。しかし、データが正しくなければ、正しい答えを導き出すことが困難となり、プロセス全体の信頼性が損なわれる。データ品質は、AI の卓越性の基礎であり、信頼できる AI の実現に寄与する。
このような背景をふまえて、AI 社会を適切に実現し、データ駆動型社会へと導くために、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持続的に確保するため、何をすべきか整理したのが本ガイドである。（第 1.01 版 2025 年 12 月公開）
https://aisi.go.jp/output/output_framework/data_quality_management_guidebook/

⑧ AI セーフティに関する具体的な影響の調査報告書

AI システムが社会に与える具体的な脅威状況を調査し、その結果を取りまとめた報告書を公開することで、「何がどのように問題になり得るか」を具体像として提示した。これまでのガイド（AI セーフティに関する評価観点ガイド）で調査してきた、AI システムが直接エンドユーザーに与える影響から、AI が社会に与える影響に範囲を拡大した調査である。AI システムの評価設計の議論を、影響ベースで進めるための基礎資料として位置づけられる。（2025 年 10 月公表）

https://aisi.go.jp/output/output_information/251031/

⑨ AI セーフティ・アプローチブックの公開（AI セーフティの普及に向けた文書）

AI セーフティインスティテュート（AISI）から公表している「AI セーフティに関する評価観点ガイド」、「AI セーフティに関するレッドチーミング手法ガイド」（以下、「両ガイド」という。）について、より多くの方に AI セーフティの実現に向けて活用していただきたいと考えている。これから両ガイドを初めて参照する方にも概要をご理解いただくため、簡潔に要点をまとめたアプローチブック（リーフレット・パワーポイント）を作成した。（2025 年 3 月公開）

https://aisi.go.jp/output/output_information/250326/

2） 制度・国際標準・適合性評価への対応

① AI 事業者ガイドライン(1.1 版)の公開

2025 年 3 月に、総務省、経済産業省が我が国における AI ガバナンスの統一的な指針である AI 事業者ガイドラインの 1.1 版を公表した。AI による新たなリスクの整理、最新技術動向や AI ガバナンスに関する国内外の動向の反映等、更新が行われた。

AI 事業者ガイドラインは、AI を活用する者が AI のリスクを正しく認識し、自主的に必要となる対策を実行できるように、開発者・提供者・利用者など様々な立場の事業者に向けた、我が国における AI ガバナンスの統一的な指針とされており、AI セーフティに関する評価観点ガイド等、AI セーフティの展開の基盤となっている。

AISI は経済産業省と共同事務局として AI 事業者ガイドライン検討会を運営し、継続的な更新に取り組んでいる。

https://www.ipa.go.jp/disc/committee/expert-group-on-aigfb.html

② ISO/IEC JTC1/SC42（AI/人工知能）国際標準関係

・国際

春季総会（インド）および秋季総会（オーストラリア）に参加した。次回総会は春にシンガポールで開催される。

主な出来事としては、金融分野における AI 標準化として共同 WG（JWG7）が設置された。AI システムの適合性評価に関するハイレベル・フレームワークを規定する ISO/IEC42007 の開発が ISO CASCO と共同（JWG6）で進められている。また、機能安全と AI や Red Teaming など含めて AI テスティング技術についても規格開発が進んでいる。日本主導で SC42 事務局セクレタリと全体の運用について意見交換を行い、JWG が多く設置されている点について JTC1 セクレタリも交えて設置基準の明確化等に関する意見交換を行った。

・国内

SC42 幹事、WG1 および JWG6 の幹事を拝命し、SC42 国内委員会、小委員会の運営に携わり、投票案件の審議、取りまとめを行った。

・省 12 プロジェクト（産総研のプロジェクト）

HMT（ヒューマン・マシン・チーミング）および AIMS に関する取組を継続し、AIMS（AI マネジメントシステム）に関係する調査として、カナダの関係者と意見交換を行った。カナダ規格協会（SCC）および EY による金融系のパイロット評価プロジェクトについて意見交換を実施した。

・その他

日本産業標準である JIS の原案作成に従事し、JIS Q42005 および JIS Q42006 の委員長として、（昨年度 42001 に続く）原案作成委員会を各 6 回開催し、参加委員の協力、日本規格協会の支援のもと原案作成を行った。

人工知能学会全国大会（大阪）において企画セッションを実施し、AISI の紹介および AI 適合性評価についてプレゼンを行った。

③ 標準型 BRIDGE 関係

「AI 分野における Joint Certification の検討」を推進した。

AI 適合性評価について、調査、検討を進めている。我が国の勝ち筋

シナリオ策定をゴールとし、認定機関および標準開発機関の専門家からなる有識者委員会を AISI 事業実証 WG の一つとして発足した適合性評価 SWG と合同で組成し既存の適合性評価制度の整理、AI 分野における適合性評価の在り方について HMT を題材に検討を行った。

調査としては、ルクセンブルク、イギリス、スイスにおいて、欧州標準化機関である CEN/CENELEC JTC21 WG2 コンビナ、SC42 英国代表、SC42JWG6 コンビナと意見交換を行った。

イギリス、ドイツ、フランスにおいて、Microsoft 関係者、アランチューリング研究所（ATI)、ドイツの認定機関である DAkkS、政府と企業とを結びつけ、政策の社会実装を行っている非営利組織 GovTech、フランス規格協会 AFNOR の関係者と意見交換を行った。Joint Certification、42007、ドイツ SC42 の取組、フランス SC42 関係者、ドイツにおける 42001 認証、AI Standard Hub の取組（3 月のイベント含む)、JTC21 議長・副議長との欧州 AI 法と整合規格の開発状況等について意見交換を行った。

京都大学と共同研究を実施し LE4SDS について検討した。法と標準（欧州 AI 法と整合規格、New Legislative Framework:NLF などを踏まえて）について検討し、法が機能していることの証跡としての適合性評価について検討した。AI など新技術に対する法のあり方（Legal Tech から Legal Engineering へ）および法執行とテクノロジーの在り方について検討した。

3) 国際連携の深化

① 「Hiroshima Global Forum for Trustworthy AI」を開催

AI セーフティ・インスティテュート（AISI、所長：村上明子）は、内閣府科学技術・イノベーション推進事務局 人工知能政策推進室と合同で、「Hiroshima Global Forum for Trustworthy AI」を 2026 年 1 月に開催した。本フォーラムは、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 AI 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する場として企画した。

https://aisi.go.jp/activity/activity_international/260130/

② AISI 国際ネットワーク（国際共同テスト演習・評価方法論の共有）

AI セーフティの技術的な内容を検討するための国際協力枠組みとして、2024 年 11 月に AISI 国際ネットワーク(International Network of AI Safety Institutes)の第 1 回会合が開催された。2025 年 2 月の

パリ会合（第 2 回）以降、共同テスト演習を軸に各国と評価手法・知見の共有を進め、2025 年 7 月のバンクーバー会合（第 3 回）では第 2 回・第 3 回演習の評価レポートが公表された。2025 年 12 月のサンディエゴ会合（第 4 回）では、ネットワーク名称を「International Network for Advanced AI Measurement, Evaluation Science」とすることが合意され、日本は AI セーフティ評価のための評価ツールの紹介を行うなど国際連携を深めている。2026 年 2 月にはインド AI インパクトサミットに合わせてインド会合（第 5 回）が実施され、ネットワークの合意文書として「AI 評価に関するコンセンサスと課題」の文書が公表された。国内では 2025 年 5 月に人工知能学会全国大会(JSAI2025)にて AISI が共同オーガナイザを務めるオーガナイズドセッション（OS-42：大規模モデルの安全対策）を開催し、AISI からは AISI 国際ネットワークや共同テスト演習の取り組みについて報告した。

関連リンク：

https://aisi.go.jp/activity/activity_international/250211/
（2025 年 2 月、第 2 回共同テスト演習レポート：ブログ記事）

https://aisi.go.jp/assets/pdf/International-Network-of-AI-Safety-Institutes-Joint-Testing-Exercise-Improving-Methodologies-for-AI-Model-Evaluations-Across-Global-Languages.pdf
（2025 年 2 月、第 2 回共同テスト演習レポート：概要版）

https://sgaisi.sg/wp-api/wp-content/uploads/2025/06/Improving-Methodologies-for-LLM-Evaluations-Across-Global-Languages-Evaluation-Report-1.pdf
（2025 年 6 月、第 2 回共同テスト演習レポート：詳細版）

https://aisi.go.jp/activity/activity_international/250718/
（2025 年 7 月、第 2 回・第 3 回共同テスト演習レポート：記事）

https://sgaisi.sg/wp-api/wp-content/uploads/2025/07/International-Joint-Testing-Exercise_3JT-Blogpost.pdf
（2025 年 7 月、第 3 回共同テスト演習レポート：概要版）

https://sgaisi.sg/wp-api/wp-content/uploads/2025/07/International-Joint-Testing-Exercise_3JT-Eval-Report-v2.pdf
（2025 年 7 月、第 3 回共同テスト演習レポート：詳細版）

https://www.gov.uk/government/news/efforts-to-share-best-

practices-on-ai-measurement-and-evaluations-driven-forward-through-the-international-network-for-advanced-ai-measurement-evalua
（2025 年 12 月、UK：ネットワーク関連発表）
https://www.aisi.gov.uk/blog/international-ai-network-consensus-and-open-questions
（2026 年 2 月、UK：AI 評価に関するコンセンサスと課題）

③ 米 Anthropic 社と MOC(Memorandum of Cooperation) を締結

AI セーフティ・インスティテュート（AISI、所長：村上明子）は、2025 年 10 月 29 日（水)、米国の AI 開発企業である Anthropic 社（共同創業者兼 CEO ダリオ・アモデイ）と MOC(Memorandum of Cooperation) を締結した。本取り組みは、信頼できる AI エコシステムのための協力に向けた覚え書きとなる。
https://aisi.go.jp/activity/activity_international/251029/

④ Singapore AI Safety Red Teaming Challenge（国際レッドチーミングへの主要参加)

シンガポールで開催される国際的なレッドチーミング・チャレンジに継続して参加している。2026 年 1 月の開催回では、AISI に加えて国内の民間セキュリティ企業等にも呼びかけ参加し、レッドチーミングの技術的な知見を得るほか、参加各国との交流を通して国際連携を図った。実践的な攻撃・評価の場で得た知見を国内外の AI セーフティの議論や取組に還元していく。
https://www.imda.gov.sg/activities/activities-catalogue/singapore-ai-safety-red-teaming-challenge

4) 産業・社会への展開

① Chief AI Officer ガイドの公開（Chief AI Officer ガイドブック及び Chief AI Officer 設置・AI ガバナンス実務マニュアル)

生成 AI やエージェント型 AI の普及により、企業活動のあらゆる場面で AI の利活用が進む一方、法令遵守、セキュリティ、プライバシー、公平性、説明責任といった複合的なリスクも同時に増幅している。こうした状況では、個別部署の工夫だけでは統制が追いつかず、全社横断で AI を「安全かつ効果的に」機能させるための、統一されたガバナンスと運用の枠組みが不可欠になりつつある。

「Chief AI Officer ガイド」は、主に民間事業者が Chief AI Officer（CAIO）を設置・運用する際の標準的な実務指針を提供し、AI による価値創出（ROI 最大化）と責任ある活用（リスク低減・規制遵守）の両立を支援することを目的としている。CAIO の役割を明確化したうえで、組織設計、プロセス、評価・監督、教育、人材、調達などの要素を統合的に示し、各社が自社の文脈に合わせて着実に AI 活用を進められるようにすることを狙いとしている。

主に CAIO 本人とその任命者を対象とした「Chief AI Officer ガイドブック」と、主に CAIO 本人とその実務スタッフ、CAIO の任命を検討する経営陣、ならびに法務・コンプライアンス・情報システム・データ・人事・事業部門などの関係者を対象とした「Chief AI Officer 設置・AI ガバナンス実務マニュアル」の二つの文書を提供する。（2026 年 3 月公開予定）

https://aisi.go.jp/output/output_information/260205/

② AISI 事業実証ワーキンググループの設置と活動および成果

AI セーフティ評価の活動を広く一般に普及させ、AI の利活用を促進させることを目的として、2025 年 3 月、AI セーフティに関する事業実証ワーキンググループ（WG）を設置した。今年度は事業実証ワーキンググループの下に、4 つのサブワーキンググループ（ロボティクス SWG、ヘルスケア SWG、データ品質 SWG、適合性評価 SWG）を組成し、30 を超える組織が参加している。

信頼とイノベーションが両立する AI 社会の実現に向けたアジェンダおよび指針として機能することを目標として 2025 年 6 月にビジョンペーパーを公開した。各 SWG では、ユースケースを選定し、評価実証を行うという流れで活動を進め、活動の成果を上期報告会、下期報告会にて報告した。

内閣府・AISI 主催の国際的なフォーラム「Hiroshima Global Forum for Trustworthy AI」において、事業実証 WG の活動を紹介した。

ロボティクス分野の AI セーフティ普及・啓発活動として動画を YouTube に公開し、国際ロボット展へ出展した。

https://aisi.go.jp/lp01/
（事業実証成果報告 LP）
https://aisi.go.jp/output/output_information/250630/
（ビジョンペーパーの公開）

https://aisi.go.jp/activity/activity_information/251003/
https://aisi.go.jp/activity/activity_information/260311/
（2025 年度 AISI 事業実証 WG 報告会）
https://aisi.go.jp/activity/activity_international/260130/
（Hiroshima Global Forum for Trustworthy AI）
https://www.youtube.com/watch?v=pKGVWA8F3p8&t=47s
（AISI ロボティクス SWG 紹介動画）

③ AI セーフティに関する活動マップとターミノロジーのチャレンジ
（活動の全体像の可視化）

国際的な文書の整理・ベンチマーキングを踏まえ、AI セーフティ実現に必要な活動を包括的にまとめた「AI セーフティに関する活動マップ」を 2025 年 2 月に公表し、見落とされがちな領域や活動間の関係を俯瞰できるようにした。公表したマップを活用した政策文書の自動分析フレームワークを 2026 年 3 月に学術論文として公表予定。
https://aisi.go.jp/output/output_information/250207_1/

④ 「LLM の安全性ベンチマークを All Japan/One Team で構築するプロジェクト」（組織横断的な連携に基づく評価手法づくり）

開発者の立場から AI の安全性評価手順を具体化し、ベンチマークとして構築・提案することを目的に、AISI 主導で議論の枠組みを立ち上げ、2025 年 10 月にキックオフを開催した。利用場面・内容・対策といった分類ごとに、2025 年度を通じて分科会で設計方針の検討やサンプルデータ整備を進めている。
https://aisi.go.jp/activity/activity_information/251009/

１．３ AISI 関連のその他の活動と成果

1） AI セキュリティ意見交換会

未だ定義が一定していない AI セキュリティの実像を明らかにしていくべく、次の三点を中心に産官学各方面の有識者から意見を聴取する場を設けている。

- AI システムのサイバーセキュリティ面からの堅牢化
- AI が組み込まれた社会システムの脆弱性や新たな脅威の把握
- AI が組み込まれた社会システムの脆弱性評価

2025 年 2 月を皮切りに 7 月、10 月、2026 年 1 月と意見交換会という形式で 4 回開催した。

第二回以降は、有識者のみならず、11 におよぶ IT 関連の業界団体も加わり、さらに広範で活発な議論が展開されている。

2) AI の動作・分析・利用方法の説明に関するアンケート調査レポート

本調査は、国内の AI 利活用実態の定点観測の視点で 2024 年から実施している。 アンケートの中では、利用者に対する AI システム提供者・AI 業務管理者などの説明の適否や AI システムの性能や信頼性、正しい利用方法などの説明責任の状況などの回答を求めた。

本調査結果は、IPA ウェブサイト内の IPA テクニカルウォッチページにて公開されている。

https://www.ipa.go.jp/security/reports/technicalwatch/20251023.html

3) 米国における AI ガバナンスとサイバーセキュリティリスク管理との統合に関する調査

生成 AI を中心とした AI 先進市場である米国における AI ガバナンスとサイバーセキュリティリスク管理との統合に関する調査を、米国調査会社と実施。具体的には、インタビューと事例収集に重点を置き、より AI 活用が進んでいる金融セクターを中心に、AI のユースケース、リスク、ベストプラクティスを調査・分析した。

本調査結果は、勉強会開催を通じて、AISI 関係府省庁へも広く展開した。

4) リソースクロスウォーク（リソース間の相互参照手順の標準化）

生成 AI の普及に伴い増える多様なリソース（ガイド、評価データセット等）について、整合性・補完性を含む相互運用性の議論を可能にするため、クロスウォーク手順を標準化して「リソースクロスウォーク」として提案する。AISI も取り組んできたガイドクロスウォークの活動について方法論を標準化していく取り組みであり、2026 年 3 月に概要を説明するディスカッションペーパーを公表予定としている。

5) 国際活動

AISI 国際ネットワークをはじめとする各国 AI セーフティ 関係者と対面やオンラインで意見交換をするとともに、講演や研修講師などを実施

している。

主な会議相手

・ASEAN、英国、欧州委員会、オーストラリア、カナダ、ケニア、大韓民国、フランス、米国、マレーシア、マルタとの意見交換
・海外 AI 関連企業との意見交換
・JICA 研修、ASEAN-Japan Cybersecurity Capacity Building Centre（AJCCBC）と JICA の共同研修

6） India AI Impact Summit

2026 年 2 月 16 日-20 日にかけて行われた India AI Impact Summit に参加。サミット本イベント、及びサイドラインで開催された”AI Safety Connect Day”での登壇に加え、関係部局との意見交換により、Japan-AISI の国際社会におけるプレゼンスを戦略的かつ多面的に向上させた。また、AISI 所長級会合では、AISI 国際ネットワークの今後の運営に向けた実務的な意見交換を実施した。

7） 国内普及活動

国内での AI セーフティに関する啓発活動として、日本経済団体連合会（経団連）、AI ガバナンス協会、東京都中小企業知的財産シンポジウム、LLM シンポジウム、AI 品質マネジメントシンポジウムなどで講演、パネルディスカッションへ参加した。

１．４ AISI の体制

1） 全体

内閣府と協力し、12 関係府省庁、5 関係機関が連携する枠組みを構築した。AISI 関係府省庁等連絡会議及び運営委員会を設置し、方針の検討を行っている。

【関係府省庁】

内閣府（科学技術・イノベーション推進事務局）、国家安全保障局、
国家サイバー統括室、警察庁、デジタル庁、総務省、
外務省、文部科学省、厚生労働省、農林水産省、経済産業省、
国土交通省、防衛省

【関係機関】

情報通信研究機構、理化学研究所、国立情報学研究所、産業技術総合研究所、情報処理推進機構

2) 日本 AI セーフティ・インスティテュートパートナーシップ

AISI の活動を効果的に推進するため、AI セーフティに関して知見を有する 5 関係機関（情報通信研究機構、理化学研究所、国立情報学研究所、産業技術総合研究所、情報処理推進機構）とパートナーシップ協定を締結し、共同で研究及び調査などを実施し、国内外への情報発信等を行っている。

2025 年 9 月に、AISI は各パートナーシップ機関における AI の安全性に係る取組を公開した。

https://aisi.go.jp/activity/activity_j-aisi_partnership/250929/

3) 事務局

事務局は IPA 内に置き、6 つのチームで構成される。

■ 戦略・企画チーム
・政府機関との調整
・民間活動との調整
・国際調整
・プロジェクト管理
・総合企画

■ フレームワークチーム
・調査
・ガイド整備
・フレームワーク的国際調整

■ 技術チーム
・調査
・アクティビティ・マップ整備
・技術ガイド整備
・技術的検証
・技術的国際調整

■ セキュリティチーム
・調査
・セキュリティ技術的国際調整

■ 標準チーム
・国際標準の動向調査（ISO/IEC JTC1/SC42（Artificial intelligence）での活動含む）
・標準的国際調整

■　利活用チーム
・事業実証ワーキンググループの企画・調整
・AI セーフティに関する AI 利活用事例等の調査

2026 年 3 月時点で常勤、非常勤含め 30 人程度である。AI セーフティに求められるスコープが広がる中、米国 AISI 約 60 人、英国 AISI 約 200 人、EU AI Office 約 60 人などと比較して、少数の専門家集団で AI セーフティの最先端の活動に取り組んでいる。

１．５　AISI の特殊な状況による課題

AISI は AI セーフティに関する技術的な専門チームとして組織しているが、利用者視点で考えれば、AI セーフティだけを深堀するのではなく、利活用も含めた方が理解しやすく、一体化した取り組みへの要望がある。また、AI 事業者ガイドラインの別添3.AI 開発者向け「適切なデータの学習」において、「データに個人情報、機密情報、著作権等の権利又は法律上保護される利益に関係した問題が生じ得る情報が含まれていないか、確認を実施」と記載されるなどデータ品質への対応も求められている。2026 年 2 月に開催されたインド AI インパクト・サミットにおける「AI 評価に関するコンセンサスと課題」の中で、「目的を明確にした評価が必要である」こと等がコンセンサスとして挙げられる一方、実際のユースケースとアプローチを検証し、AI 測定・評価の在り方が課題として挙げられるなど、国際的な議論も政策面、技術面の両面でより進展している。AI セーフティとして技術だけでは解決しない分野への対応を関係者と検討する必要がある。

AISI は、対府省庁、技術者、海外と幅広いステークホルダーへの対応が求められる中、国内のパートナーと連携し、AI 分野の人材確保に向けた積極的な取り組みを進めている。しかし、国内外で AI 人材が不足する現状において、公務員の待遇制約も影響し、厳しい状況である。

また、関係機関が多いうえ、バーチャル組織のため異なる情報管理ポリシーや環境下での運用も求められ、ガバナンス体制が重層的で、臨機応変な機動的な対応が難しい。

さらに、OECD は 2024 年 11 月に公表した「Assessing potential future Artificial Intelligence risks, benefits and policy imperatives」で将来の AI リスクの可能性として 10 のリスクを挙げている。そこでは技術的リスクのほかに「ガバナンスメカニズムや組織が、速く変化する AI の改革についていけなくなる」というリスクを指摘している。

２今後に向けて

２．１　考慮すべき新たな要素

1.5 の課題に加え、今後に向けて考慮すべき要素には以下のものがある。

1)　法規及び基本計画

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 日に公布され、2025 年 9 月 1 日に全面施行された。また、AI 法第１８条に基づく人工知能(AI)基本計画(2025 年 12 月 23 日閣議決定)の中では、関連技術の研究開発及び活用の推進に関する施策として「AI ガバナンスの主導」が掲げられ、人員を直ちに現行の２倍程度に拡充する等の AISI の抜本的強化、AI モデルの技術的評価、AI がもたらすリスクに係る実態把握について述べられている。

2)　技術的な進展

急速に発展を見せる AI エージェントや、ロボット等に実装されるフィジカル AI など、AI の利活用のすそ野が広がるとともに、AI システムの発展が加速度的に進行している。これらの新たな AI 及び AI の用途拡大により、新たな AI セーフティの課題が顕在化する可能性がある。また、AI セーフティのスコープ、観点、評価手法が変化する可能性もある。

3)　他国の AI セーフティ関係機関

I. AISI 国際ネットワークの取り組み

AISI 国際ネットワークは、2025 年 2 月のパリ AI アクション・サミット及び 2026 年 2 月のインド AI インパクト・サミットでの議論を踏まえ、情報交換から具体的なベストプラクティスの共有へと発展しつつある。また、AISI 国際ネットワーク内でバイでの技術的連携を模索する動きもある。

II. 国際機関などによる取り組み

G7、および OECD や国連などの国際機関でも AI に関する議論が進められており、その検討内容や推進体制、具体的な方法が国内外に影響を与えると考えられる。その議論は主に AI の倫理的利用や規制の在り方、国際協力の強化といったテーマを中心に展開されており、日本を含む各国の政策や体制に大きな影響を及ぼす可能性がある。

4） 民間企業との連携

英国や米国の AISI は、民間企業と連携しながら、AI モデルの評価や評価ツールの開発に取り組んでいる。また、これらの評価ツールは産業界や社会全体での実用性を考慮し、長期的な影響やリスクにも対応できるよう、戦略的な視点で設計されている。米国政府と民間企業との連携状況を注視しながら、日本も AI セーフティ評価を行うための連携を模索する必要がある。

２．２　今後の取り組み方向

現状見えている AI を取り巻く課題と考慮すべき新たな要素を踏まえ、以下の取り組みを通じて、安全・安心で信頼できる AI の提供に必要な情報や技術を AISI が有し、日本を世界で最も AI を開発・活用できる国にすることを目的として、必要な対策を講じる。

1） AI 安全性についての評価指標を作る

AI の安全性の確保を促進するガイドについて、進化する最新 AI に対応した既存ガイドの改訂を行うとともに、個別の産業分野特有の課題に特化したガイドの作成を行っていく。また、産業分野毎のベンチマーク開発、データセットの充実化を図るとともに、我が国初の国際標準の策定を推進していく。

2） 自ら評価する能力をもつ

AI の第三者評価が可能となる評価環境の構築を本格化し、構築した AI 評価環境によるフロンティア AI の悪用評価に繋げていく。また、AI によるサイバー攻撃・防御への対応として、AI インシデントに係る情報の収集及び評価を行い、政府関係機関と情報共有する。

3） AI 関連情報を収集・分析・提供する

官民が連携した業界毎の AI ガイド、制度設計への技術的助言を行うとともに、影響が大きな分野（電力網などインフラ分野・モビリティ）への AI 導入事例や、AI インシデント事例につき幅広に情報収集・分析を行う。

また、AI 法 16 条調査等への技術的支援（ディープフェイク、著作権侵害等）や、リスクの高い AI 製品の流通を防止するための技術的情報を関係者に提供する。

4)　国際協調を主導する

上記 1)～3)の活動に資するため、AISI 国際ネットワーク間の連携を強化し、ベストプラクティスの共有に積極的に参画する。また、新興国のキャパシティビルディングへの技術的協力、国際機関の連携も行っていき、懸念される AI の評価に関し日本 AISI が国際的な対応をリードできるよう取り組む。