2026年5月21日
JAI-Trust：日本の生成AIの安全性とセキュリティーの
ベンチマーク構築プロジェクト

主催者挨拶

AIセーフティ・インスティテュート所長

村上　明子

AISI Japan
AI Safety Institute

AIセーフティ・インスティテュート（AISI：エイシー）

AISI Japan AI Safety Institute

安全性と信頼性を専門的かつ中立的な立場で検証する「公的な第三者機関」AIセーフティ・インスティテュート（AISI）

日本のAISI（エイシー）の概要

名　称

（日本語）AIセーフティ・インスティテュート
（英　語）Japan AI Safety Institute（略称 J-AISI）

業務内容

- 安全性評価に係る調査、基準等の検討
- 安全性評価の実施手法に関する検討
- 他国の関係機関（英米のAI Safety Institute等）との国際連携に関する業務

関係機関

内閣府、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、文科省、厚労省、農水省、経産省、国交省、防衛省

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

主要な実績

AIセーフティに関する評価観点ガイド、レッドチーミング手法ガイド等の公開。AISI国際ネットワーク(米・EU等の主要国AISI関連機関10カ国が参加)**に参加**し、AIの共同テストに関するトラックをリード。

村上 明子
(AIセーフティ・インスティテュート所長)
(SOMPOホールディングス株式会社執行役員常務　グループChief Data Officer)

2024年、AIの安全性と信頼性を専門的かつ中立な立場で検証する公的な第三者機関であるAIセーフティ・インスティテュート(AISI)の設立に伴い所長に就任。

AISIの役割とスコープ

AISI Japan AI Safety Institute

「信頼できるAI」の提供に必要な情報や技術を有し、
日本を世界で最もAIを開発・活用しやすい国にすることがAISIの目的

役割

主に3つの役割を担う。

●AI安全性に関連する情報のハブとして、信頼できる優れたAIを活用した製品やサービスの普及

●AI技術やAI安全性等に係る国際標準化等により、我が国のAI関連分野のイノベーションの加速・競争力強化

●情報収集・分析・共有・対策の早期実施により、AIの開発・普及に伴う国民の生命・財産に危害が及ぶような事象など、国家の安全を脅かす事象の抑止

AIの開発や利用をする者が**AIのリスクを正しく認識**できる仕組みの構築 ＋ ガバナンス確保などの必要となる対策を**ライフサイクル全体で実行**できる仕組みの構築 ⇔ **国内・国際的な関係機関**

▼

イノベーションの促進と**ライフサイクルにわたるリスクの緩和**を**両立**する枠組みを実現

スコープ

- AIによる以下の事象や検討事項の中で、諸外国や国内の動向も見ながら柔軟にスコープを設定し取組を進めていく。

社会への影響	ガバナンス	AIシステム	コンテンツ	データ

AISI強化策について

AISI Japan AI Safety Institute

設立から2年が経過し、AISIのミッションの見直しを実施し、
規模も30人*から60人規模へ拡大予定。 *2026年3月時点

- 新たなミッションとして、政府関係機関として、AIの安全性に関する調査・情報提供などを行うことを通じ、以下の観点から貢献することを検討。
  - AI安全性に関連する情報のハブとして、**信頼できる優れたAI活用製品・サービスの普及**に貢献する（もって、国民生活の向上・社会課題の解決を促進する）
  - 我が国の**AI関連分野のイノベーションの加速・競争力強化**のため、我が国主導の国際標準化等に技術の側面から貢献する
  - AIの開発・普及によって、**国民の生命・財産に危害が及ぶような事象**など、国家安全を脅かす事象の抑止に技術の側面から貢献する

ソフト・ロー(自主的対応の促進)によってAI導入の障壁を取り除き、民間事業者のAIによる価値創出と責任ある活用の両立を支援する。そのために、

（１）AIの物差しを作る
（２）自ら評価する能力を持つ
（３）AI関連情報を収集し、分析し、提供する
（４）国際協調する 役割を担う

AISI業務におけるベンチマークプロジェクトの位置づけ

AISI Japan AI Safety Institute

AISIにおいてAI安全性を評価する環境を構築するにあたり、
本プロジェクトで構築する「**AI安全性ベンチマーク**」が礎となる。

評価指標を作る

評価観点ガイドの策定
及び評価ツールの公開

・評価観点をLLM以外にも拡充
・ベンチマークの整備

評価観点ガイド
&
AI安全性ベンチマーク

評価する

安全性を高める為の
評価環境を整備・評価

・評価環境の基盤構築
・適合性評価制度の具体化

評価環境

LLMシステム

AIエージェント

Physical AI

信頼できるAIの開発による
利活用の促進

「安全」で「安心」なAIを提供するためには

AISI Japan AI Safety Institute

「**安全**」とは「許容不可なリスクがないこと」
「**安心**」とは「心配・不安がないこと」

- 「**安全**」はISO/IEC GUIDE 51:2014(E)で定義されている
  - Safety: Freedom from risk which is not tolerable
- 「**安心**」とは「心配・不安がなくて、心が安らぐこと。また、安らかなこと（広辞苑）」
- 主観的な要素が強い。

「AIを安全に使う技術的な手段を提供する」だけではなく
「安心してAIを使えることができるために必要な情報を提供する」
ことも必要

AISI

Japan AI Safety Institute