

ヒアリング

「AI技術の安全性とAISIIの取り組みについて」

AISII 村上

AIセーフティ・インスティテュート（AISI）について

- AIが社会に与える影響が拡大するなかで、その**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」**である**AIセーフティ・インスティテュート（AISI）**が必要不可欠に。国際的にAIガバナンスの重要性が高まる中、**AI安全性サミット（2023年11月、英国）**を契機に「**AI安全性**」を具現化するための議論が進み、英・米はAISIを設置。
- 我が国も第7回AI戦略会議（2023年12月）における**岸田元総理からの指示を踏まえ**、2024年2月に日本の**AIセーフティ・インスティテュート（所長：村上明子氏）**を設置。
- AI安全性の知見のハブとして、**国内外の関係機関とのネットワークを強化中**。**AISIの安全性の評価能力を 確立しながら、安全性評価のための基準、ガイダンスを作成**。



日本の^{エイシー}AISI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。

（※）米、豪、加、EU、仏、日、ケニア、韓、シンガポール、英



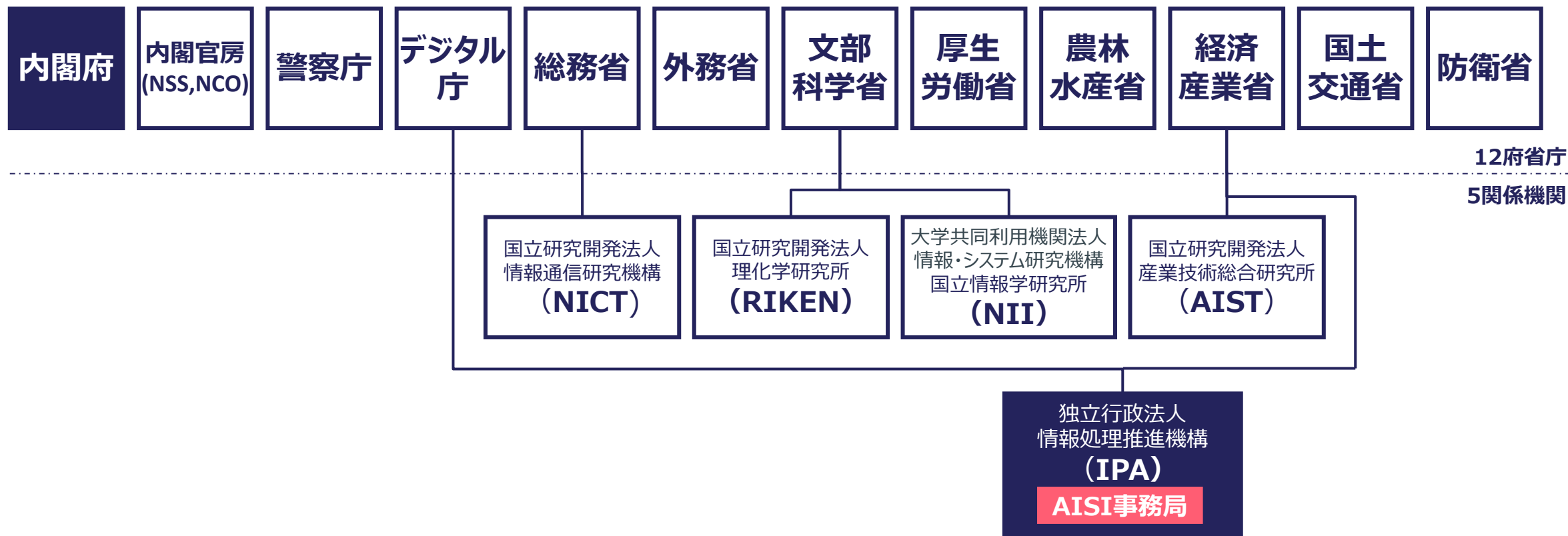
村上 明子

**（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的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は、12府省庁・5関係機関が横断的に参画する**政府関係機関**
事務局は経済産業省とデジタル庁を所管官庁としている**IPA**内に設置

* 2025年4月時点



「AISIのパーパス：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・活用できる国にする」

● パーパス実現のためAI基本計画等を踏まえて、ミッション（AISIの役割・責任）を整理

- ① 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活用製品やサービスの普及**を促進する（もって、国民生活の向上・社会課題の解決を促進）
- ② 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**を支える
- ③ 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**に貢献する

● ミッション達成に必要なAISIが行うべき4つの活動

- (1) AIの**安全性に関する**物差し（**ベンチマーク**）を作る
- (2) **AIモデル・システム**を自ら評価する能力を持つ
- (3) AI関連情報の収集・分析・提供を**率先して行う**
- (4) **AI安全性に関して**国際協調を主導し、**国際競争力**を高める

AISIの国際連携

国境を越えて展開するAIの安全性については、国際協調が必須。下記3つの枠組みを通じて、国際プレゼンスの向上と国際協調を進める。

- 1) AISI国際ネットワーク(NAAMES/ネームス)を通じた同志国連携
- 2) 二国間協力
- 3) 国連等の多国間枠組みでの情報収集と発信

米国の呼びかけで10カ国が参加し、現在は英国がコーディネーターを担当し、AISI国際ネットワーク(NAAMES/ネームス)を形成。技術フォーラム、所長級会合、バイ会合等で連携。



AISI国際ネットワーク初会合



AISI国際ネットワーク
所長級会合



内閣府・日本AISI共催
国際フォーラム

NAAMES以外にも多数の国際連携を行い、AI安全性に係る最新の取組にコミット



欧州AI安全保障会議



国連ハイレベルウィークの
独立国際科学パネル



日韓AISI技術ワークショップ

AISIとフロンティアAI企業との連携

1. アンソロピック社との協力覚書（MoC）締結（2025年10月）

- 米国アンソロピック社のダリオ・アモデイCEOの初来日に伴い、高市総理への表敬はじめ、日本政府と米・アンソロピック社との間で、信頼できるAIエコシステムについて意見交換を実施。
- このMOCは、**AI評価の科学を進展させる**ことを目的とし、**AIモデルの能力、限界、潜在的リスクを企業が評価するための手法**に関する情報交換とベストプラクティスの共有を行うことへの**合意**である。
- AI技術の進展や将来の動向に関しても**継続的に意見交換**を進め、国際的に適用可能な共通基準の検討を支援し、**責任あるイノベーションを加速**させる。



出典：令和7年10月29日 アモデイ・アンソロピック社CEOによる表敬 | 総理の一日 | 首相官邸ホームページ

2. OpenAI社との協力覚書（MoC）締結(2026年5月)

- 米国OpenAI社のジェイソン・クウォンCSOが来日し、**AISIとの間で**、AIの安全性評価等の協力に係る覚書（MOC）の署名を実施。
- このMOCは、AIの**安全性評価**および**ベンチマーク、サイバーセキュリティ**に関する協力の可能性を探ることを目的とし、業界別の安全性評価や評価手法に関する知見、ベストプラクティスの共有について意見交換を行うことへ**合意**である。
- **AIの安全性評価やサイバーセキュリティに関する継続的な対話**を通じて、国際的に適用可能な**安全性評価の在り方やベンチマークの課題**について検討を進める。



出典：ジェイソン・クウォンCSOの大臣表敬訪問の様子 - 内閣府