資料2

AIセーフティ・インスティテュート（AISI）の今後の活動について

2024年２月

内閣府 科学技術・イノベーション推進事務局
IPA　AIセーフティ・インスティテュート

内閣府 Cabinet Office　AISI Japan AI Safety Institute

AISIの設立について

内閣府 Cabinet Office　AISI Japan AI Safety Institute

概　要

AIの安全性に対する国際的な関心の高まりを踏まえ、AIの安全性の評価手法の検討等を行う機関として、米国や英国と同様に、日本においても、AIセーフティ・インスティテュートを２月14日に設立した。

同機関は、内閣府をはじめ関係省庁、関係機関の協力の下、IPA（独立行政法人情報処理推進機構）に設置され、諸外国の機関とも連携して、AIの安全性評価に関する基準や手法の検討等を進めていく。

所長には、元日本IBMのAI研究者で、現在は損保ジャパンCDO（チーフ・デジタル・オフィサー）で京都大学防災研究所客員講師の村上明子氏が就任した。

業務内容（暫定）

1. 安全性評価に係る調査、基準等の検討
   - ①安全性に係る標準、チェックツール、偽情報対策技術、AIとサイバーセキュリティに関する調査
   - ②安全性に係る基準、ガイダンス等の検討
   - ③上記に関するAIのテスト環境の検討
2. 安全性評価の実施手法に関する検討
3. 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

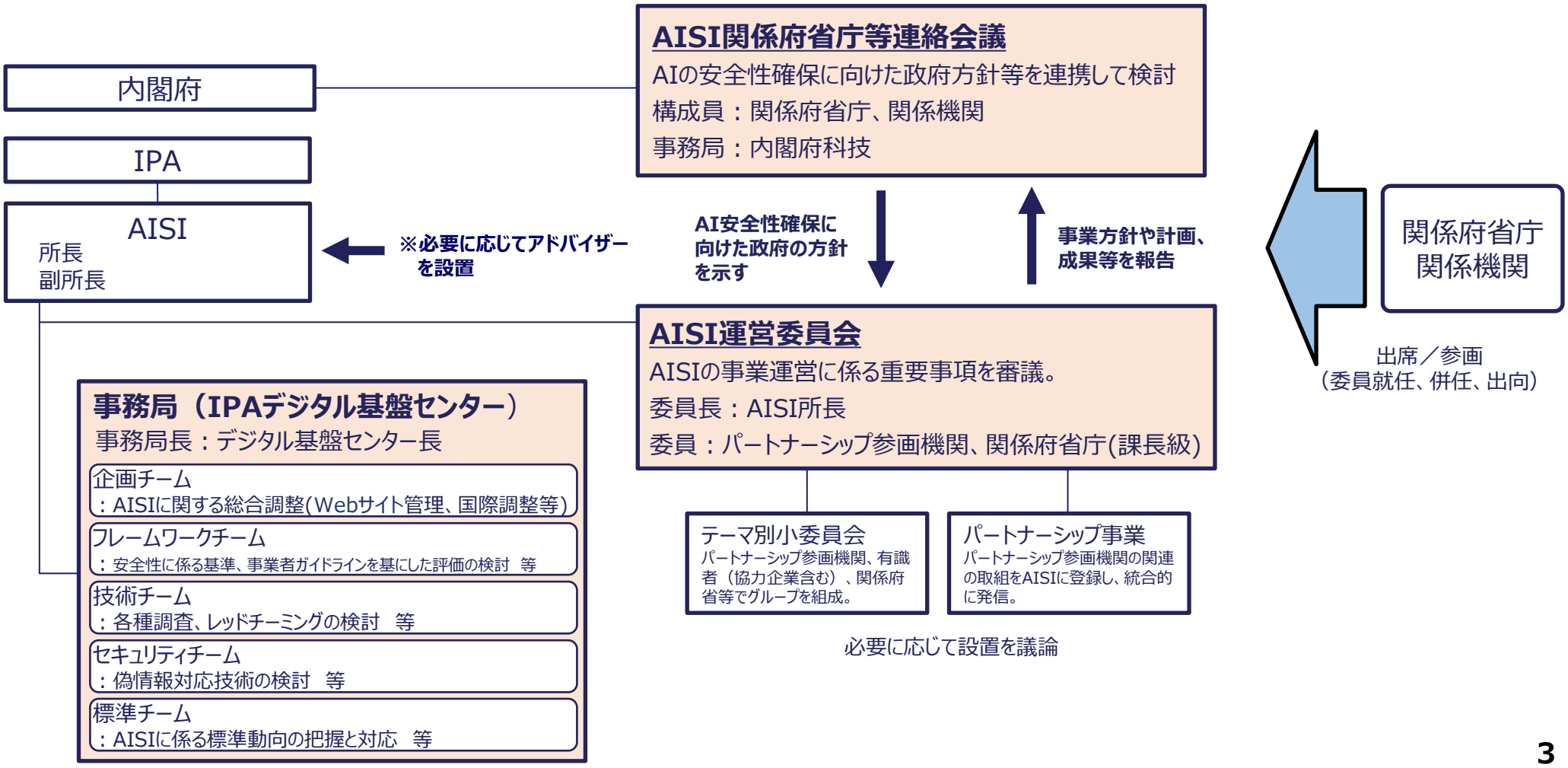
関係省庁・関係機関

関係省庁　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、内閣サイバーセキュリティセンター、警察庁、デジタル庁、総務省、外務省、文科省、経産省、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

AISIの体制（案）

- 内閣府を事務局とする「AISI関係府省庁等連絡会議」を設置し、重要事項を審議（年間２～３回の開催を予定）。
- AISIの中に、AISI所長を委員長とする「AISI運営委員会」を設置（月１回の開催を予定）。運営委員会の下に、必要に応じて、「テーマ別小委員会」や「パートナーシップ事業」（研究機関等の関連の取組みをAISI事業として発信）を設置。
- AISIの事務局として、IPAデジタル基盤センターの中で５つのチームを編成。



AISIの業務とAISI事務局チームの対応

AISI業務内容	AISI事務局チームの対応	備考（現時点での予定）
1．安全性評価に係る調査、基準等の検討		
①安全性に係る標準、チェックツール、偽情報対策技術、AIとサイバーセキュリティに関する調査	**・技術チーム ・セキュリティチーム （・標準チーム）**	**・４月目途に各種調査事業の方針を定める**
②安全性に係る基準、ガイダンス等の検討	**・フレームワークチーム ・標準チーム （・セキュリティチーム）**	**・３月末目途にリスクマネジメントフレームワーク(RMF)の和訳を公表する ・５月末目途にAI事業者ガイドラインと RMFとのCrosswalkを公表する**
③上記に関するAIのテスト環境の検討	**・技術チーム**	**・８月目途にレッドチーミングテストの手順（案）を策定**
2．安全性評価の実施手法に関する検討		
産学との意見交換 AI安全性評価の運用に係る検討 （上記１で作成する基準等の活用）	**・企画チーム ・フレームワークチーム**	**・７月目途に安全性の評価観点を整理**
3．他国の関係機関との国際連携に関する業務		
海外の関係機関との連携（※） 付随する基礎調査　など	**・企画チーム**	**・３月に英米AISI所長や国内関係機関等との意見交換を実施。 ・５月までに今後の国際協力方針のセットを目指す。**

（※）他国の関係機関との国際連携に関する業務について、我が国政府の取組との整合性が適切に確保されるよう、AISIは関係省庁に対して、対応方針等について事前に照会／情報共有するとともに、結果等について速やかに情報共有する。

【参考】英国AIセーフティ・インスティテュートの設立

- 2023年11月2日、**スナク首相は、AI安全性サミットの機会をとらえ、「AIセーフティ・インスティテュート」の計画**を発表
- 「AIセーフティ・インスティテュート」は、**米国、シンガポールやGoogle DeepMindといった国・企業とも連携しながら、AIの安全な開発について国際協力を促進**
- 「AIセーフティ・インスティテュート」は、これまで暫定的にスナク首相直轄で活動してきた「フロンティアAIタスクフォース」を発展的に改組し、常設の機関とするもの

AIセーフティ・インスティテュートの役割

1. 先進的AIシステムについての評価手法を確立するとともに、評価を実施

商業的な圧力から独立した形で政府が外部評価を実施するとともに、標準化・ベストプラクティスの促進を支援。外部評価の重点項目は、①デュアルユース能力を有するAI、②社会的インパクトの大きなAI、③システムの安全性とセキュリティ、④統御不可能性（人間の思いもよらない挙動）

2. AI安全性に関する基礎的研究を実施

研究重点項目は、①**AIガバナンスのためのツールの開発**、②評価技術の向上、③より安全なAIシステムに関する新たなアプローチ

3. 情報交換の促進

英国政府に加え、米国、シンガポール等各国とAIの安全性に関する情報交換を行うとともに、AI産業界、アカデミア等とも連携

我々の**AIセーフティ・インスティテュートはAI安全性のグローバルなハブ**として活動し、進展の速い技術の能力・リスクについて重要な研究を行い、世界をリードする（中略）あらゆる人々に裨益する**AIの安全性確保について、パートナー国やAI関連企業から支援を得られることは喜ばしい。**これは長期的な英国の国益にとって正しいアプローチ

【参考】米国AIセーフティ・インスティテュート・コンソーシアムの立上げ

内閣府 Cabinet Office　AISI Japan AI Safety Institute

- 2024年2月8日、**レモンド商務長官**は、国立標準・技術研究所に設けられる**AIセーフティ・インスティテュートにAIセーフティ・インスティテュート・コンソーシアム（AISIC、AI Safety Institute Consortium）を立ち上げる**ことを発表（**2月7日には、AIセーフティ・インスティテュートのリーダー**も発表）
- コンソーシアムには、AIに携わる**200以上の幅広いステークホルダー（大企業・スタートアップを含む企業、市民社会、アカデミア、注）が参加**
- コンソーシアムは**５つのワーキンググループ**に分かれ、AI安全性に関してテーマ別に取り組む

記者会見に臨むレモンド商務長官

本日、私たちは米国AIセーフティ・インスティテュート・コンソーシアムを立ち上げた。これは、**安心、安全で信頼できるAIを前に進めるため、米国の200以上の先進的なAIステークホルダーを結集**したもの。私たちはこの（重要な）瞬間に合わせる用意はできている（Xから引用）

米国は、**AIのリスクを低減し、大きな潜在的な能力を強化するため、標準を策定し、ツールを開発するにあたり、大きな役割**を有している。バイデン大統領は2つの目標を達成するためにあらゆる手段を活用するように指示している。すなわち、**安全性基準を定めることと、イノベーションのエコシステムを保護すること**である。これこそ米国AIセーフティ・インスティテュートに期待すること（商務省プレスリリースから抜粋）

AISICのワーキング・グループ

①生成AIのリスクマネジメント：AIリスクマネジメント・フレームワーク（AI RMF）に補完的なリソースの開発、AI RMFの実用化

②合成コンテンツ：合成コンテンツに関する既存の標準、ツール、手段及び実践の特定、科学的な知見に基づく標準、技術の検討

③能力評価：AIが損害を引き起こす可能性のある能力に焦点を当てたAIの能力（注）評価及び監査のガイダンス、基準の作成、AI技術の開発を支援するための、テストベッドなどのテスト環境の利用可能性の確保

（注）化学、生物学、放射線、核、サイバーセキュリティ、自律的複製、物理システムの制御など

④レッドチーミング：AI開発者、特にデュアルユースの基盤モデル開発者がレッドチーミングテストを実施し、安全で安全性の高い信頼性のあるシステムの展開を可能にするために適切なガイドライン、手続き、プロセスの確立

⑤安全性・セキュリティ：デュアルユースの基盤モデルの安全性とセキュリティの管理に関するガイドラインの調整及び開発