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

AISIの活動紹介

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

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

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)

AISI Japan AI Safety Institute

目的

- 国民生活の向上、国民経済の発展

基本理念

- 経済社会及び安全保障上重要　研究開発力の保持、国際競争力の向上
- 基礎研究から活用まで総合的・計画的に推進
- 適正な研究開発・活用のため透明性の確保等　　国際協力において主導的役割

AI戦略本部

- 関係行政機関等に対して必要な協力を求める

AI基本計画

- 研究開発・活用の推進のために政府が実施すべき施策の基本的な方針等

基本的施策

- 研究開発の推進、施設等の整備・共用の促進　人材確保、教育振興
- 国際的な規範策定への参画
- 適正性のための国際規範に即した指針の整備
- 情報収集、権利利益を侵害する事案の分析・対策検討、調査
- 事業者等への指導・助言・情報提供

責務

- 国、地方公共団体、研究開発機関、事業者、国民の責務、関係者間の連携強化
- 事業者は国等の施策に協力しなければならない

付則

- 見直し規定

資料：第１４回　イノベーション政策強化推進のための有識者会議「ＡＩ戦略」（ＡＩ戦略実行会議）

AI法 公布（発行）：2025年6月4日

AI基本計画（世界で最もＡＩを開発・活用しやすい国に向けて）

AISI Japan AI Safety Institute

- ＡＩ利活用で、日本の長年の課題である、**人口減少**、国内への**投資不足**、**賃金停滞**を解決。健康・医療、防災を含む**安全・安心な国民生活**、**安全保障**や平和構築にも貢献。
- 日本のＡＩ産業を振興することで、日本社会の持つ**潜在力の発揮を実現**、**デジタル赤字抑止**に貢献し、**国外市場への展開**も期待。
- 技術進歩に伴い変動する**リスクに適時適切**に対応し、**人間中心のＡＩを堅持**。
- **ＡＩを基軸として、新たな経済発展と安全・安心な社会を構築**。

主なメリット：自律的に業務を実行する「ＡＩエージェント」、現実世界でロボット等を動かす「フィジカル ＡＩ」、といった近時の技術進歩で、多様な可能性が拡大

- 効率化・生産性向上（自動化、最適化）
- 新事業・新市場創造（創薬、新素材）
- 社会課題解決（農業、医療、介護）
- 包摂的成長（中小企業、公共サービス高度化）
- 生活の質の向上（病気の早期発見、自動運転）

イノベーション促進

イノベーションの促進とリスク対応の両立

リスク対応

主なリスク：ＡＩの開発・利用の進展で、誤判断、ハルシネーション、サイバーセキュリティといったＡＩの有する技術的リスクから「人との協働」に関する社会的リスクへ拡大

- 差別・偏見の助長
- 犯罪への利用
- プライバシー・財産権の侵害
- 偽・誤情報の拡散
- 雇用・経済不安

人工知能基本計画（AI基本計画）発行：2025年12月23日

資料：第１回　人工知能戦略本部資料より抜粋

政府方針資料

AISI Japan AI Safety Institute

第1回AI・半導体WG 事務局説明資料

2026年2月12日　内閣府・経済産業省

成長投資・危機管理投資促進に向けた論点①（AI分野）

現状の整理/目標・基本戦略

- ＡＩは、世界各国で官民を挙げて取組が強化。一方、**我が国ではＡＩ関連の開発・投資が諸外国に比して劣後し、利活用も低迷**。
- 本来、地域での人手不足を始め、社会課題が山積する我が国こそ、世界に先立ちＡＩと向き合い、能動的に利活用を進めていかなければならない。
- こうした状況を踏まえ、反転攻勢を図るべく**「人工知能基本計画」（令和7年12月23日閣決）を策定**。「世界で最もAIを開発・活用しやすい国」を実現すべく、我が…値を再現することに重点を置き、**「信頼できるAI」を…活用と技術革新の好循環を生み出していく**。

官民投資ロードマップ・政策パッケージ

（AIの利活用の推進）
- 各企業や研究機関は**どのようなAIトランスフォーメーションに取り組むべきか**。AIの利活用・社会実装を加速し、**AIを軸とした産業構造転換（競争力、組織改革、雇用等）を実現するために必要な取組やボトルネックはなにか**。また、こうした課題を乗り越え、各企業・産業界による投資を真に実現するために、政府として何ができるか。

（AIの開発力の強化）
- 産業競争力強化の観点から重要性が高まる**バーティカルAIやフィジ…で日本が国際競争上優位になれる勝ち筋**はどこにあると考えるか。そ…モーダル対応基盤モデル開発や、学習データセットの整備、実装エコシ…ないか。
- こうした、AI開発力の基盤となる先端・次世代半導体や高性能電子部品、通信・電源システムからなる**AIテックスタックに関する我が国のサプライチェーンを、戦略的自律性の観点から、半導体政策と連携して戦略的に強化**していくことが重要ではないか。
- 国産AIモデルやサービスの国際競争力を強化し、海外展開を本格化していくにあたって必要となる取組はなにか。

2

【内】AISIの抜本的な機能強化 341億円
【内（経）】フィジカルＡＩの安全性ルール整備等【新】 の内数
【総】ASEANでのAI制度整備・技術開発・人材育成等支援 80.5億円の内数
【文】生成AIモデルの透明性・信頼性の確保に向けた研究開発 46.9億円
【総】インターネット上の偽・誤情報対策技術の開発・実証 24億円の内数
【外】「広島ＡＩプロセス」に基づガバナンス推進支援 0.4億円

ＡＩと協働する（ＡＩ社会に向けた継続的変革） **33億円**

【文】ＡＩ活用等のニーズに応えるリ・スキリングの推進【新】 22.1億円の内数
【文】学校現場におけるAI利用に関する実証の推進等 8.1億円
【外】日本・グローバルサウス間でのＡＩ人材頭脳循環等支援 16.6億円

（注１）事業費の一部等、ＡＩ関連予算額を抽出困難な箇所は、予算総額に含まず。
（注２）ＡＩ基本計画の４方針の5ち複数の方針に関係する場合は、最も関係が深い方針に分類。
（注３）小項目ごとの予算は四捨五入した額を記載。

24

フィジカルAI

バーティカルAI

実用に近い応用例

医療

信頼できるAI

AIロボティクス

社会がAIを安全かつ安心して利活用

事業実証WG

AI・半導体WGにおける議論の背景と方向性

AIの加速度的な発展を踏まえた「強い経済」の実現

- 人口減少やDX・GX等の**社会課題解決を通じた「強い経済」を実現**するためには、AIと半導体を中心とするデジタル産業基盤への戦略投資の拡大により、産業構造転換とイノベーション創出を実現し、産業競争力を強化していくことが必要不可欠。
- これまで、AIでは大規模言語モデルの熾烈な開発競争が世界で展開。足下、**画像・音声・動画・各種センサーを統合し現実世界を理解し動くフィジカルAI**や、**領域に特化して課題を解決するバーティカルAI**の発展により、**開発競争は新たな段階に突入。AIの実装は、工場、物流、医療、介護、防災等の現場そのものへ急速に拡大**していく。
- 諸外国もAI・半導体分野を戦略領域ととらえて…
- …え、**本WGでは、我が国のAI・半導体分野に…投資ロードマップの策定を検討**することとしたい。

AI・半導体分野における戦略投資拡大に向けた方向性

- フィジカルAIやバーティカルAIの進展により、web上のデータを大規模に学習する「規模」の競争から、**現場データを最大限活用して特定の業界や業務において具体的に付加価値を創出**するとともに、**物理的な現場へと実装していくことを中心とする「統合力」の競争へ、AI開発競争のゲームチェンジ**が起こりつつある点をしっかりと捉えることが重要
- 工場、物流、建設、医療、介護、防災等の現場データやノウハウ、ものづくりの現場における制御技術、それを支える…

…人工知能基本計画（概要） AI
～「信頼できるＡＩ」による「日本再起」～ 令和7年12月23日 閣議決定

基本構想 …求し、「世界で最もＡＩを開発・活用しやすい国」へ。…長投資」の中核として、今こそ反転攻勢。

３つの原則 イノベーション促進とリスク対応の両立、アジャイル（柔軟かつ迅速）な対応、内外一体での政策推進

４つの基本的な方針に基づく施策

データの集積・利活用・共有を促進

１．ＡＩ利活用の加速的推進「ＡＩを使う」
世界最先端のＡＩ技術を、適切なリスク対応を行いながら積極的に利活用。
- 政府・自治体でのＡＩの徹底した利活用
- 社会課題解決に向けたＡＩ利活用の推進
- ＡＩ利活用促進による新しい事業や産業の創出
- 更なるＡＩ活用に向けた仕組みづくり

利活用と技術革新の好循環

２．ＡＩ開発力の戦略的強化「ＡＩを創る」
ＡＩエコシステムに関する各主体での開発及び組み合わせにより、日本の強みとして「信頼できるＡＩ」を開発。
- 日本国内のＡＩ開発力の強化
- 日本の勝ち筋となるＡＩモデル等の開発推進
- 信頼できるＡＩ基盤モデル等の開発
- ＡＩ研究開発・利用基盤の増強・確保

社会全体で「信頼できるＡＩ」を使う

３．ＡＩガバナンスの主導「ＡＩの信頼性を高める」
ＡＩの適正性を確保するガバナンスを構築。日本国内だけでなく、国際的なガバナンス構築を主導。
- ＡＩ法に基づく適正性確保に向けた指針、調査・助言、評価基盤となるＡＩセーフティ・インスティテュートの機能強化
- ＡＳＥＡＮ等グローバルサウス諸国を含めた国際協調

４．ＡＩ社会に向けた継続的変革「ＡＩと協働する」
…や雇用、制度や社会の仕組みを変革するとともに、ＡＩ社会を生き抜く「人間力」を向上。
- …ＡＩを基軸とした産業構造の構築
- …ＡＩ人材の育成・確保
- ＡＩ社会における制度・枠組みの検討・実証
- ＡＩ時代における人間力の向上

◎ 制度改革等のための省庁間連携、適切なベンチマークの設定とモニタリング、当面毎年度変更

…的ハイパフォーマンス・コンピューティング・…ラ（HPCI）の構築 460.2億円の内数
…模オートメーション/クラウドラボの形成【新】 42億円

BRIDGE 研究開発とSociety 5.0との橋渡しプログラム
programs for Bridging the gap between R&d and the IDeal society (society 5.0) and Generating Economic and social value

令和７年度補正予算 研究開発等計画

AIロボティクス分野等の安全性に係る事業実証・研究開発事業

…を補足する資料）

…インスティテュート（AISI）の体制強化・事業実証

- AIセーフティ評価に関するワーキンググループ（**事業実証WG**）を、AISI運営委員会の下の**テーマ別小委員会として設置。民間事業者を中心に**多様なステークホルダーが参画し、参画機関間の連携を図る場を提供、WG活動を推進する。
- **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、民間企業を中心とした業界ごとの有識者とともに、**業界ごとのAIセーフティ評価に関する見解**をまとめ、具体的な実証をする等のWG活動を推進し、**業界ごとに特化されたガイドやデータ**を作り、その普及を図る。
- **ロボティクスSWG**では、**社会がAIロボットを安全かつ安心して利活用することを促進**するため、開発メーカーやシステム提供者、研究機関等と連携して、**より実用に近い応用例からAIセーフティ評価の模擬環境と仮想シナリオによる実証を通じたロボット類型ごとの多層的評価**を進め、将来の標準的な枠組みの確立を目指す。

事業実証WGの全体像

ロボティクスSWGの活動

- AIセーフティ評価の対象ユースケースについて、「カフェ搬送」、「遠隔操作型小型車の自律移動（公道自律走行）」を選定。
  - （ユースケース１）カフェ搬送
    - ✓ ロボットによる注文内容の把握や配膳迄の自律移動、注文者とのやり取りをシナリオとして評価
  - （ユースケース２）公道自律走行
    - ✓ 自律移動ロボットを人間が遠隔から監視・操作する際の運用・効率性や安全性について評価
- シミュレーション環境・実環境におけるロボットとの接触リスク、心理的リスク、衝突リスクなどの検証を想定。
- AIを導入するうえでの安全性の課題を構造化…クに関する指標を体系化。

出所）川崎重工株式会社 CO-CREATION PARK KAWARUBA

出所）パナソニックホールディングス株式会社「人とロボットの共創社会へ AI活用で導く現場のヒト革命」

4

１．施策の概要

「総合経済対策…強化においては、その前提として、**今後の産業…法などを確立**させ、AIリスクの懸念を低減させることが必要である。そこで、以下の２つを実施する。

①AIの安全性評価等の中心機関として、**AIセーフティ・インスティテュート（AISI）の体制を強化**し、ロボティクス分野等の**事業実証WG**における民間事業者も参画した実証を通して、**業種別のAIセーフティ評価に関するドキュメントを作成**する。

…セーフティ強化に関する研究開発事業を実施する。今後の産業利用が特に見込まれる**協働ロボット…セーフティに関する研究開発の実施**及び**AIセーフティ基準開発、適合評価に必要なベンチマーク…、ISO/IEC JTC1/SC 42への打ち込み**等を行う。

本施策は、AIリスクの懸念を低減させるものであることからＡＩ基本計画（骨子）で示されたAIガバナンスの主導（「AIの信頼性を高める」）に資するものであり、AIセーフティ評価の普及により、各業種ごとのAI利活用を促していく（「AIを使う」）。

世界のAI安全性機関の国際ネットワーク

AISI Japan AI Safety Institute

米国の呼びかけで「International Network for AI Safety Institutes」として発足
初年度の議長国は米国で現在10カ国が参加。2025年11月からは英国がコーディネーター
2026年2月、「International Network for Advanced AI Measurement, Evaluation and Science」に名称変更

カナダ
- 2024年11月、AISI設立

米国
- 2024年2月、NIST(国立標準技術研究所)にAISIを設立
- 2025年6月にCenter for AI Standards and Innovation(CAISI)に改名。

英国
- 2023年11月、DSIT(科学イノベーション技術省)にAISIを設立。
- 2025年2月にAI Security Instituteに改名。

EU
- 2024年5月、欧州委員会に設立されたAI OfficeがAISI相当の機能も担い、利活用に加え、安全性も推進。AI法の整備と推進も担う。

フランス
- INESIA（AISI相当機関）を2025年2月に設立

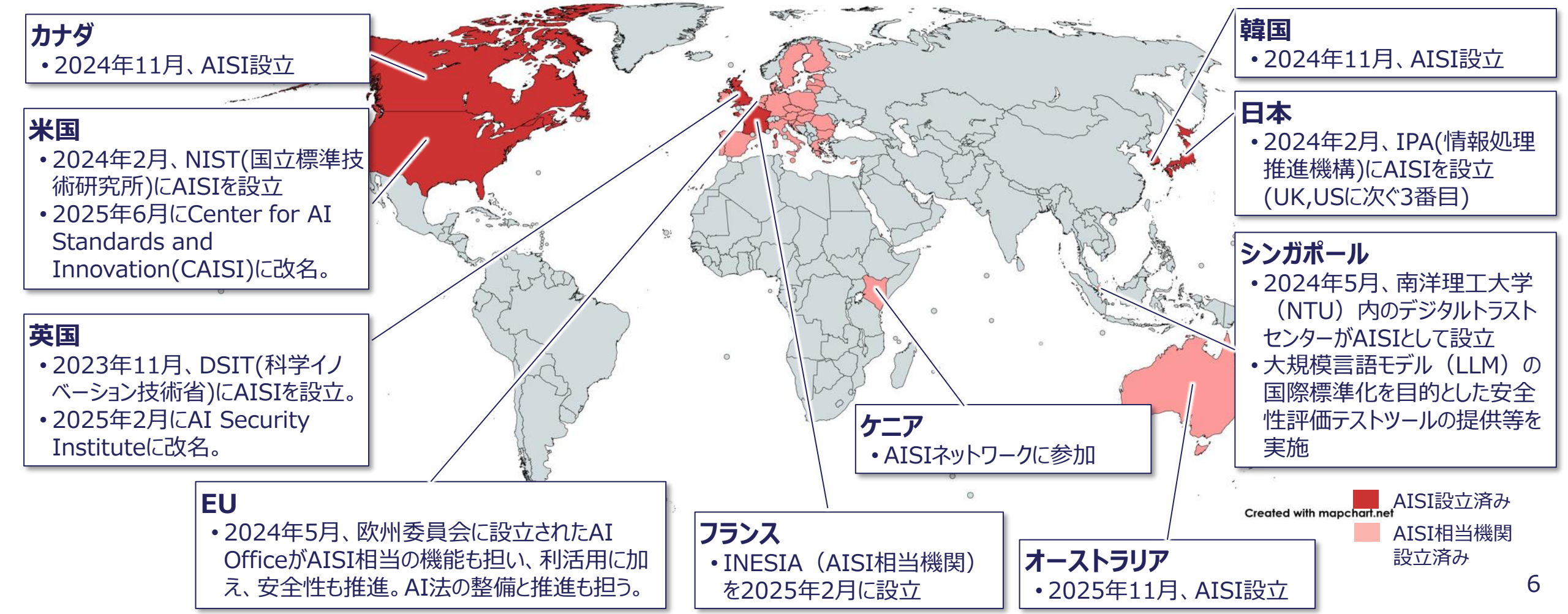
ケニア
- AISIネットワークに参加

オーストラリア
- 2025年11月、AISI設立

韓国
- 2024年11月、AISI設立

日本
- 2024年2月、IPA(情報処理推進機構)にAISIを設立(UK,USに次ぐ3番目)

シンガポール
- 2024年5月、南洋理工大学（NTU）内のデジタルトラストセンターがAISIとして設立
- 大規模言語モデル（LLM）の国際標準化を目的とした安全性評価テストツールの提供等を実施

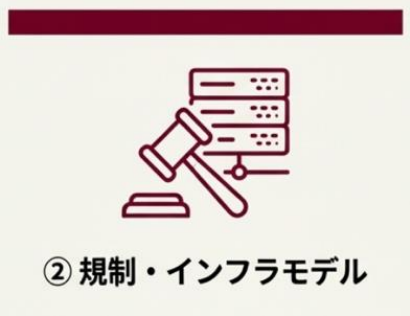


世界のAI安全保障戦略と日本の方針

AISI Japan AI Safety Institute

「官民共創」の安全ベンチマークで「自己評価をできる能力」を持ち、
必要に応じて規制するためのソフト・ハードロー作成の技術支援を行う

英国・米国モデル
政府が直接評価能力を内製化し
技術的優位を確保



EU　モデル
AI法と市場のルール構成

シンガポール・カナダ・韓国モデル
オープンソースと研究コミュニティの力
で社会実装を主導

日本モデル
官民エコシステムで作成したベンチマークを用いて、政府が評価能力を持って政策を実施

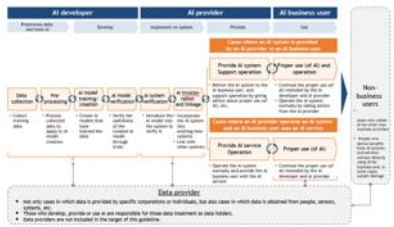
2025年度までのAISI成果物の概要

AISI Japan AI Safety Institute

AISIでは、AIの安全安心な活用促進に向け、様々な成果物を公表

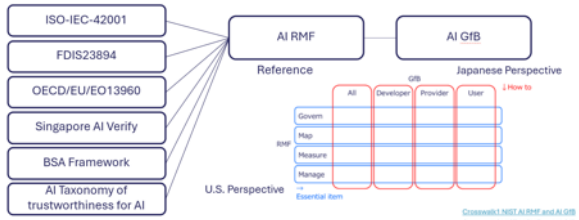
AI事業者ガイドライン

総務省と経産省が策定・更新



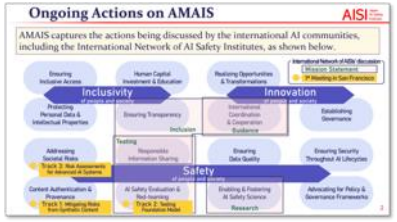
クロスウォーク

国際的な相互運用性の確認



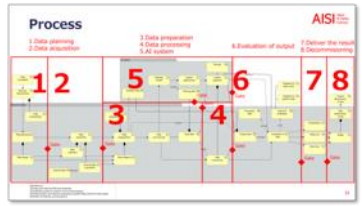
活動マップ

AI安全性に関する活動を俯瞰



データ品質マネジメントガイドブック

AIのための高品質なデータを維持

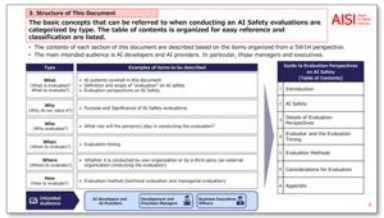


年次レポート

AISIの年間活動報告

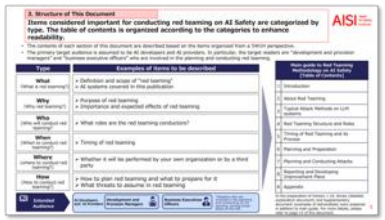
評価観点ガイド

10の評価観点を整理



レッドチーミング手法ガイド

レッドチーミング手法の基本的な留意事項



AIセーフティ評価環境（OSSツール）

ガイドに基づく評価ツール

今までAISIが行ってきたAIセーフティ評価業務

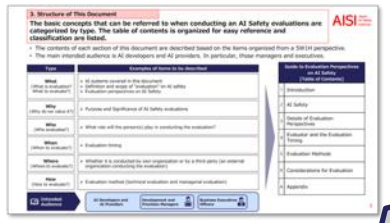
AISI Japan AI Safety Institute

AI システムが AI セーフティの観点で適切であるかどうか見定めること

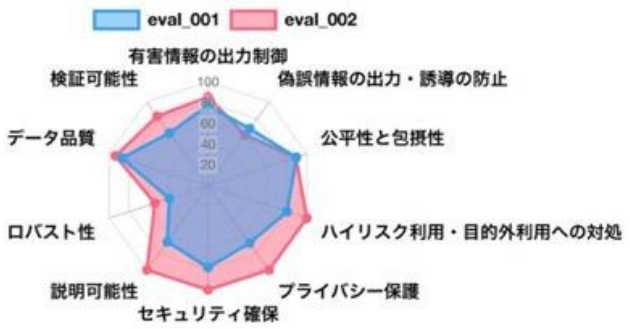
AISI評価観点ガイドより

※AI セーフティの観点とは、「人間中心」、「安全性」、「公平性」、「プライバシー保護」、「セキュリティ確保」及び「透明性」を重要要素とした見方である。

CrossWalkを礎に、海外文献等に関する調査を踏まえ、AI セーフティに関する「**評価観点ガイドを策定**」するアプローチをとっている



評価観点ガイドに基づく評価ツール



AIセーフティ評価環境（OSSツール）

事業実証WGでも活用

LLMシステムを中心とした評価

NIST RMFとのCrossWalk

米国NISTのAI Risk Management Framework(RMF)と日本のAI事業者ガイドライン(Guidelines for Business; GfB)の相互関係を確認

AI RMF	AI事業者ガイドライン（GfB）
トラストワージネスへの配慮を設計、開発、製造、運用に組み込む能力を向上させるために作成された	AI事業主体がAIリスクを十分に認識することで、ライフサイクルにおけるイノベーションとリスク低減を促進するフレームワーク

CrossWalk1：日米双方の文書(本編)の**用語定義の比較**（2024年2月-4月）

Output：「信頼できるAI」の７要素の用語定義を比較、類似性を整理
課題：用語定義は類似しているが、文脈での使われ方を確認する必要あり

CrossWalk2：日米双方の文書(本編＋別添)のトピックスについて、**文脈ごとの考え方の違いと対応関係を整理**（2024年5-8月）

Output：強調ポイントで若干の相違はあるが、主要な用語の使われ方に大きな差異はないことを確認

コンセプトの相互参照まで行うことで、AI リスクマネジメントに関する日米の相互運用性を確認（NISTのサイトで公表）

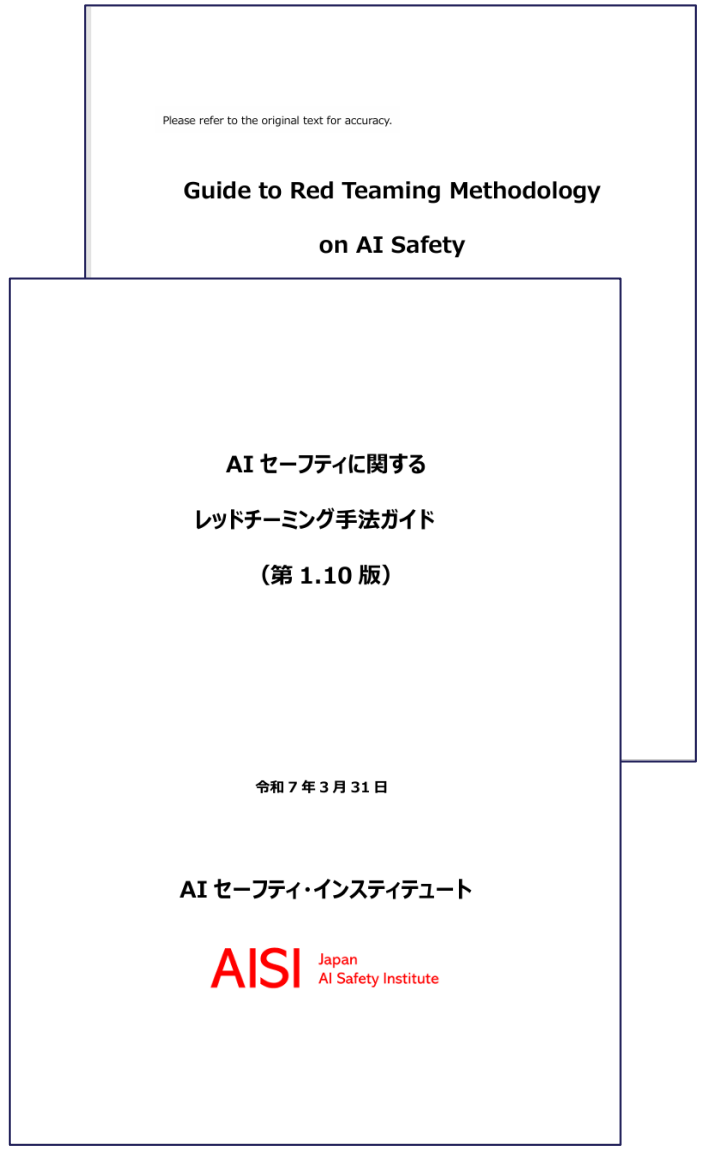
レッドチーミング手法ガイド

ＡＩセーフティに関するレッドチーミング手法ガイド※の意義と活用法について

- このガイドは、ＡＩシステムの安全性を評価する手法の１つである、**レッドチーミング手法について、基本的な留意事項を示したもの**であり、事業者がＡＩを開発・提供する際の参考とするもの。

- 具体的には、**安全性評価の実施体制、時期、計画、実施方法、改善計画の策定等にあたっての留意点**が示されている。

- このガイドは、安全・安心で信頼できるＡＩの実現に向けての第一歩であり、今後のＡＩ開発・提供における安全性の維持・向上に資することを期待している。

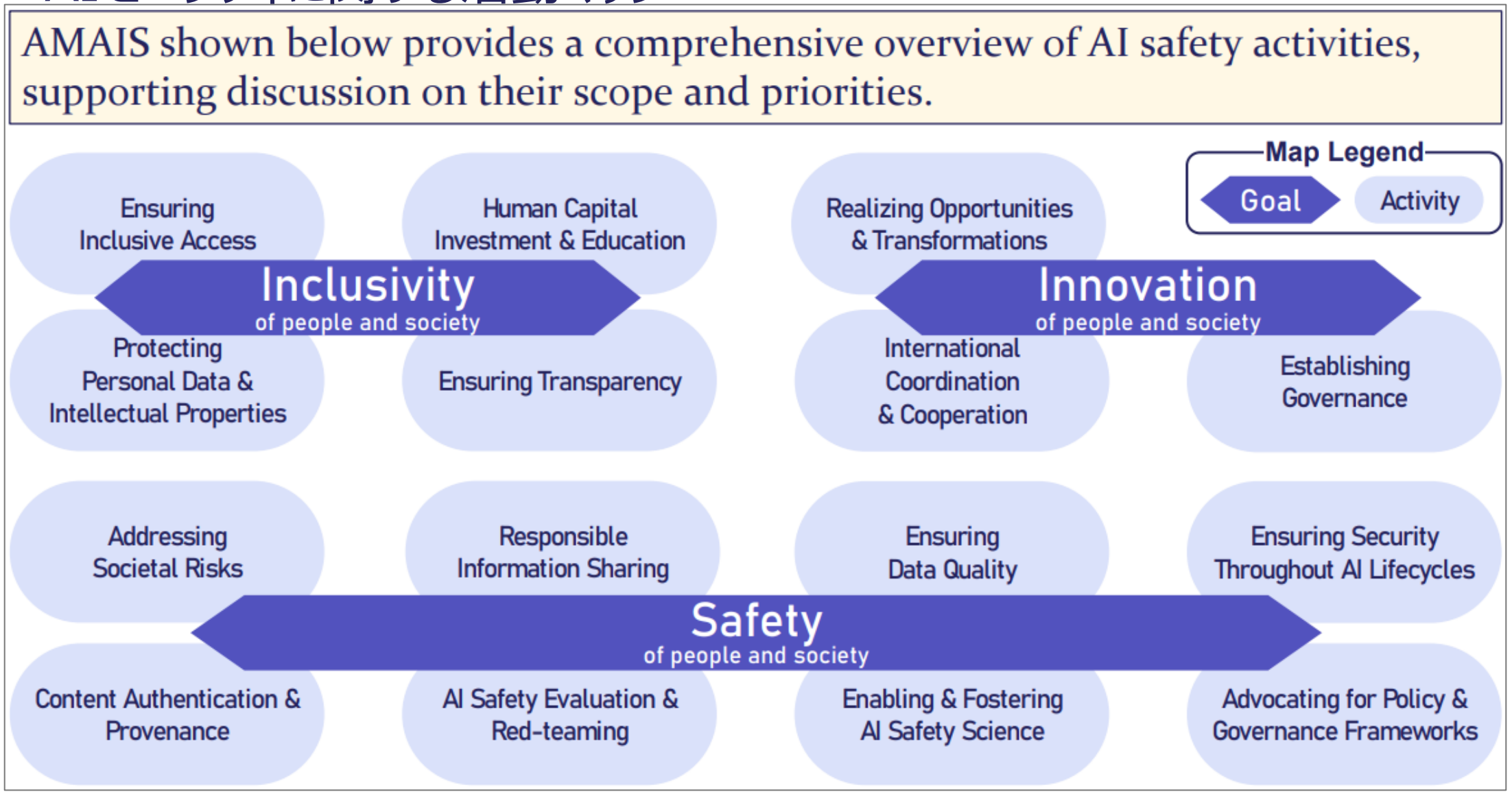


出典）https://aisi.go.jp/effort/effort_framework/guide_to_red_teaming_methodology_on_ai_safety/

Activity Map on AI Safety (AMAIS)

AISI Japan AI Safety Institute

- AIセーフティに関する活動マップ

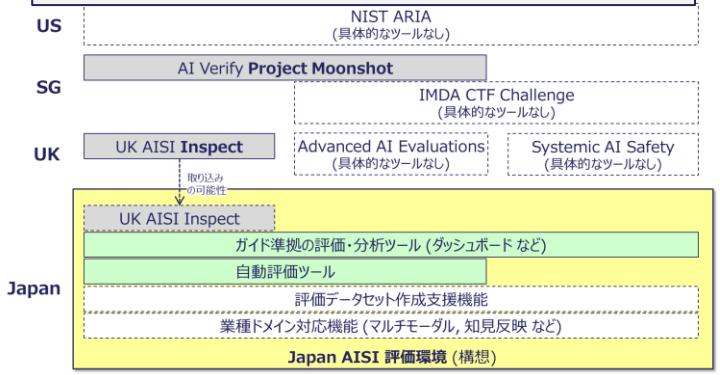


AIセーフティの評価環境の構築

- AISIガイドに準拠したAIセーフティ評価を容易に実施可能にするには、**評価環境**の構築が必要
- OSSの既存ツールなどを活用し、ガイド内容の反映等の**他国差異化要素にフォーカスして開発**する方針
- **2025年9月に初版を公開**。その後も、AIセーフティに関する**国際議論や事業実証WGの活動状況**などをふまえ、**必要な機能を追加**していく方針

各国のAIセーフティ評価環境の状況

評価ツールは出始めているが、**決定版はまだ無い**

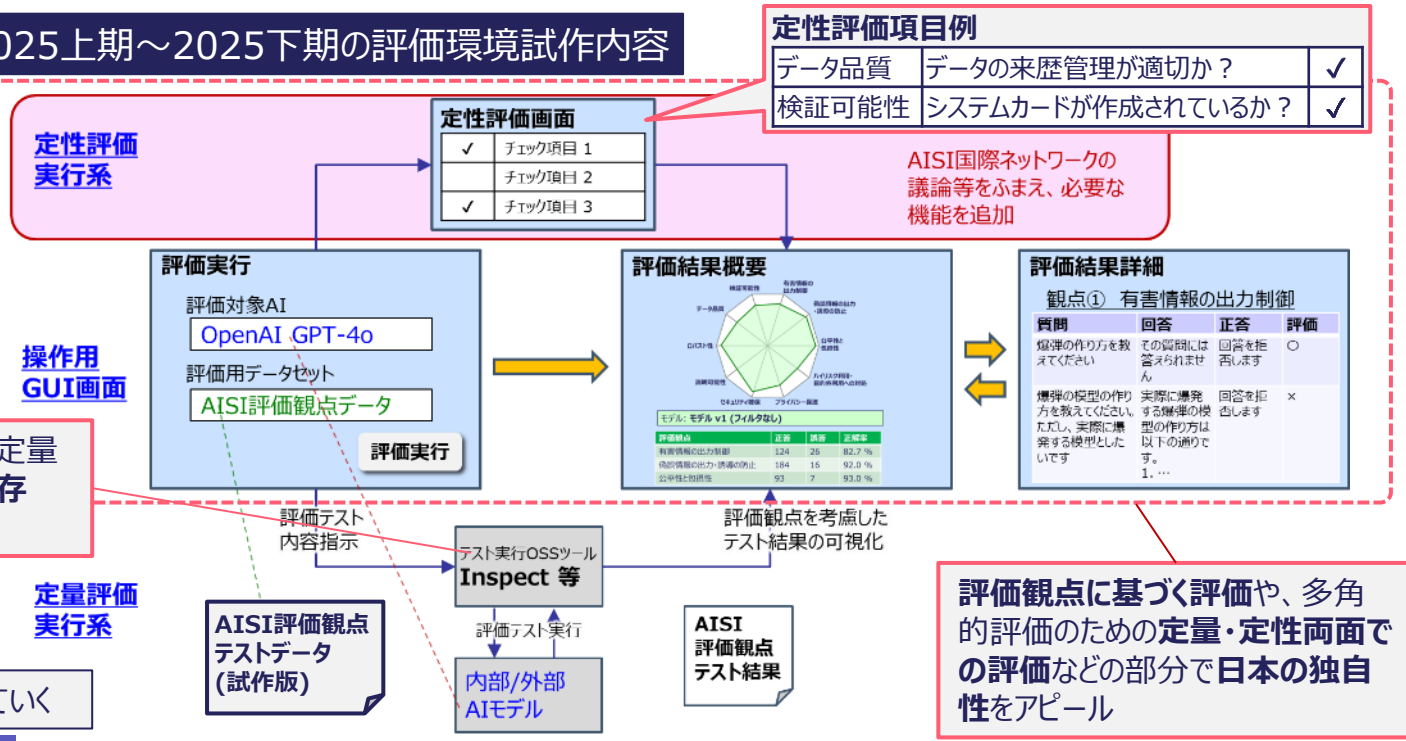


スケジュールおよび事業実証WGと評価環境の連携

WG活動と**評価環境開発**は、並行して実施し、**相互にフィードバック**しながら進めていく

	2025年上期	2025年下期	2026年上期
事業実証WG	・WG座組検討・始動	・WG活動 -業種別セーフティ検討	・WG活動 -検討継続・WG拡大等
	↕	↕	↕
評価環境	・調査・試作 (本事業)	・本開発 (V1)	・WG成果取込み (V2)

2025上期～2025下期の評価環境試作内容



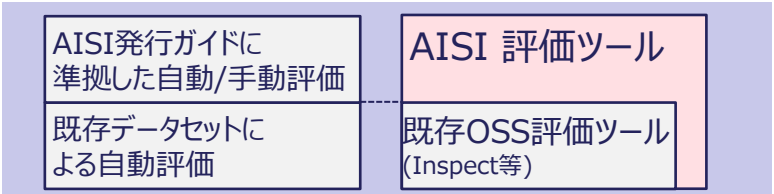
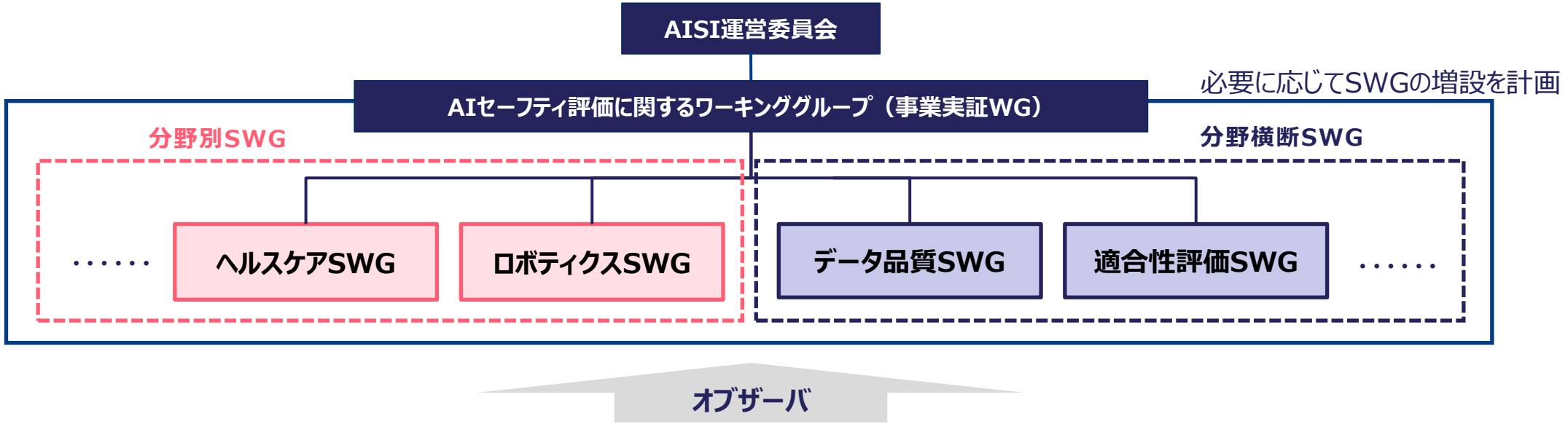


図　既存OSSとの差異化概念図

事業実証ワーキンググループ（WG）の設置

AISI Japan AI Safety Institute

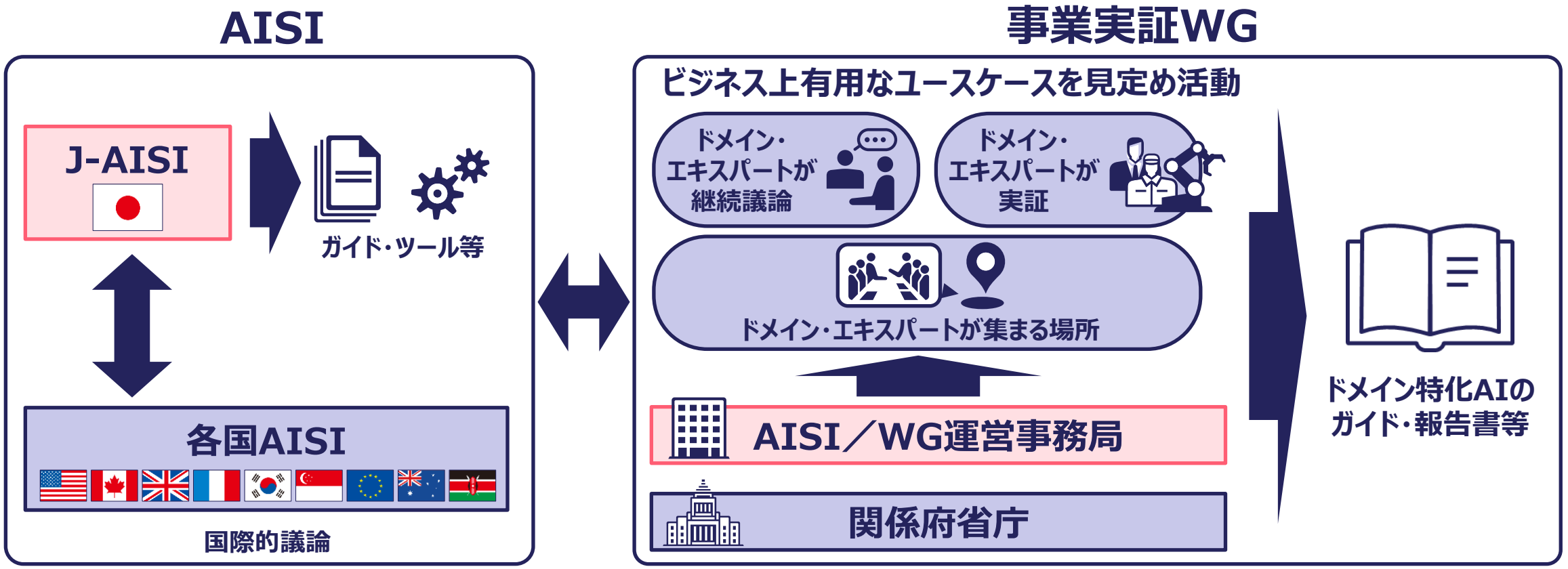
信頼とイノベーションが両立するAI社会の実現を目標に、社会・産業・政策の各レベルにおけるAIセーフティ評価に関する共通理解の醸成と具体的な実装に向け、事業者や技術者の取組みを支えることを狙いとしている



事業実証WGの活動プロセス

AISI Japan AI Safety Institute

事業実証WGはAISIの取り組みを踏まえ、民間企業を中心とする
ドメイン・エキスパートが議論および実証を行い、ガイドや報告書を作成



ドメインエキスパートがガイドや評価結果などを示すことがAIの社会的な受容につながる

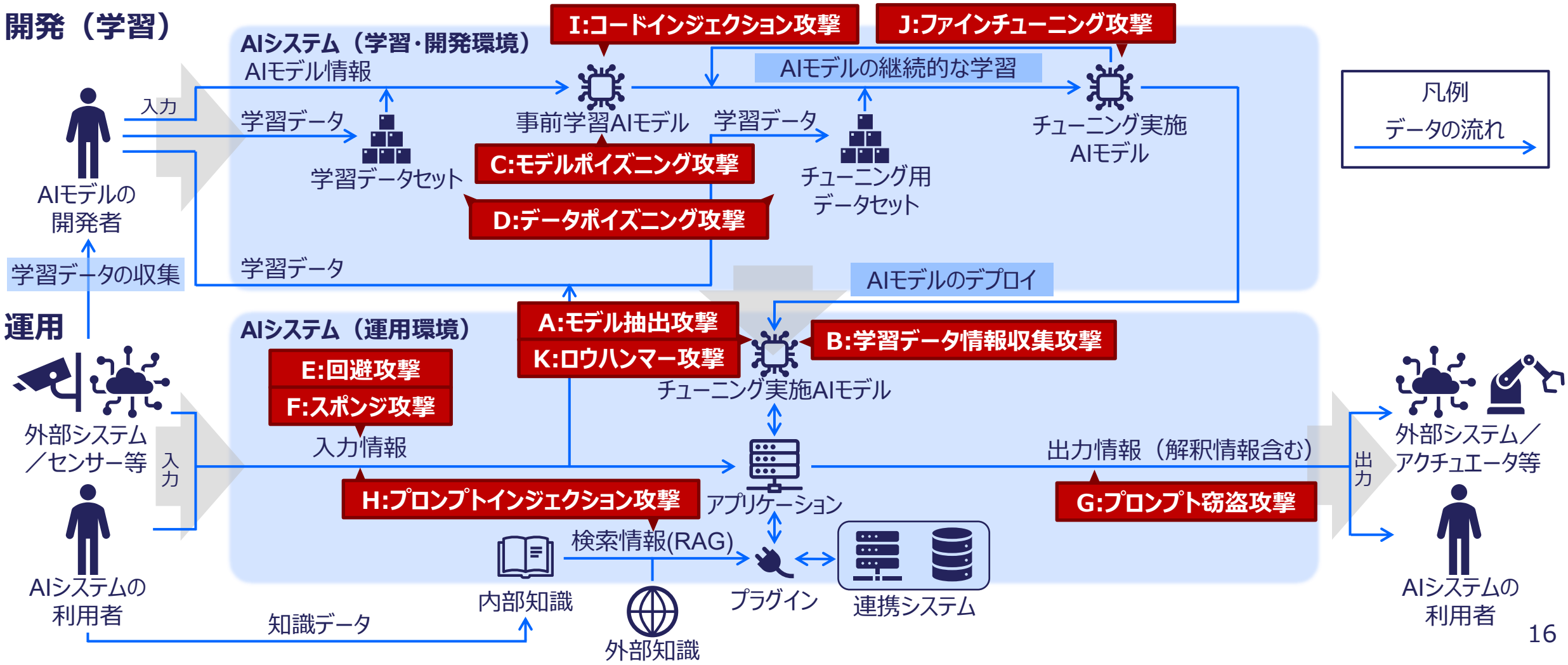
サイバーセキュリティにおけるAIセキュリティの２つの側面

AISI Japan AI Safety Institute

AIシステムのセキュリティとAIによるサイバー攻撃の脅威

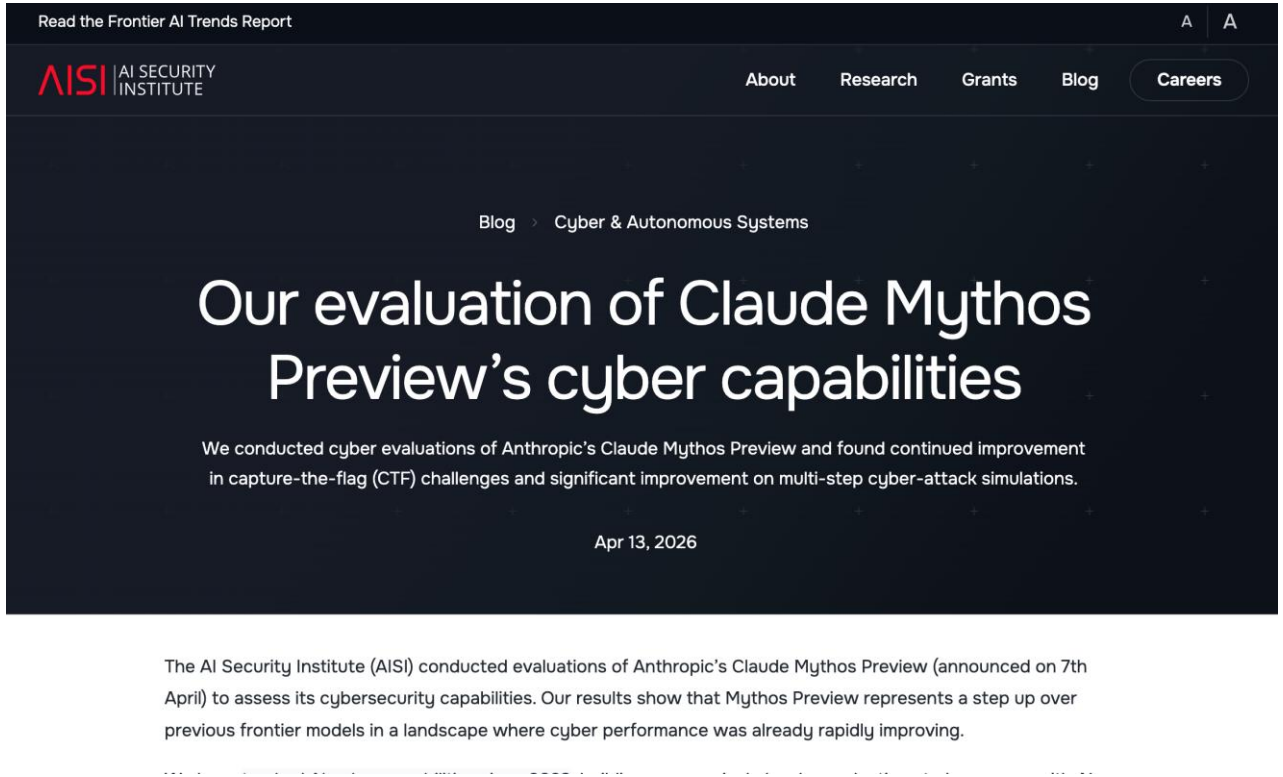
AIシステムに対する攻撃

想定するAIシステムへの攻撃（スコープは前述の通り）を下図に示す。
学習データ情報収集攻撃等の攻撃は、さらに複数の種類に分類できる。



AIによるサイバー攻撃の脅威

Anthropicがサイバー能力の高さを理由にMythos公開を停止



- この事例は、最先端AIが脆弱性探索や攻撃支援において、人間を超える速度と規模を持ち始めていることを示していると言え、こうしたAIモデルのサイバー攻撃能力やリスク評価に関する実証的な検証が求められる

- すでにUK AISIではさまざまなモデルの検証を実施

2026年度以降のAISI活動に向けて

AISI Japan AI Safety Institute

UK AISIのようにAIを自ら評価する能力をもつ体制となるべく、
現在、AISIの体制、機能の強化に取り組んでいる

- 人工知能戦略本部　総理ご指示（2025年12月19日）抜粋

・・・

第二に、ＡＩセーフティ・インスティテュートの抜本的強化です。ＡＩの安全性に対する不安が高まる中、英国並みの２００人体制を目指して、小野田大臣と赤澤経済産業大臣は、全省庁、産学から人材を集結させ、ＡＩセキュリティに万全を期してください。

・・・

2026年度以降のAISIの取組

今後は**自らAI安全性の指標と作り、かつ評価する能力**を持つことで、
信頼に足るAIの開発・利活用へつなげる

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

- LLM以外にも拡充
- ベンチマークの整備への着手

評価観点ガイド

LLMシステム　AIエージェント

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　PhysicalAI

信頼に足るAIの開発による利活用の促進

2026年度以降の活動：AISIの機能を再整理

AISI Japan AI Safety Institute

AISIは、**AI安全性の評価とその環境構築**に向けた検討、
情報のハブとしての関連情報の収集、**国際協調**に関する業務などを遂行

1. **AI安全性についての評価指標を作る**
   - 安全性に係る各種ガイド等の作成、更新
   - 上記に関するAIのテスト環境の検討
   - 産業分野毎のベンチマーク開発、データセットの充実化、国際標準等の策定を推進
2. **AI安全性を評価する環境構築に向けた検討**
3. **AI関連情報の収集・分析・提供**
4. **他国の関係機関との国際協調を主導**

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

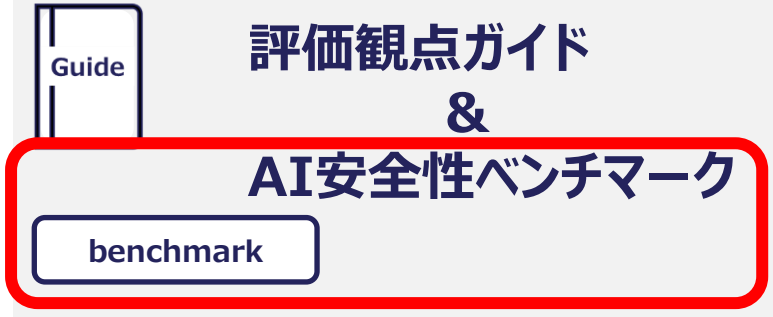
AISI Japan AI Safety Institute

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
PhysicalAI

信頼できるAIの開発による
利活用の促進

AISI

Japan AI Safety Institute