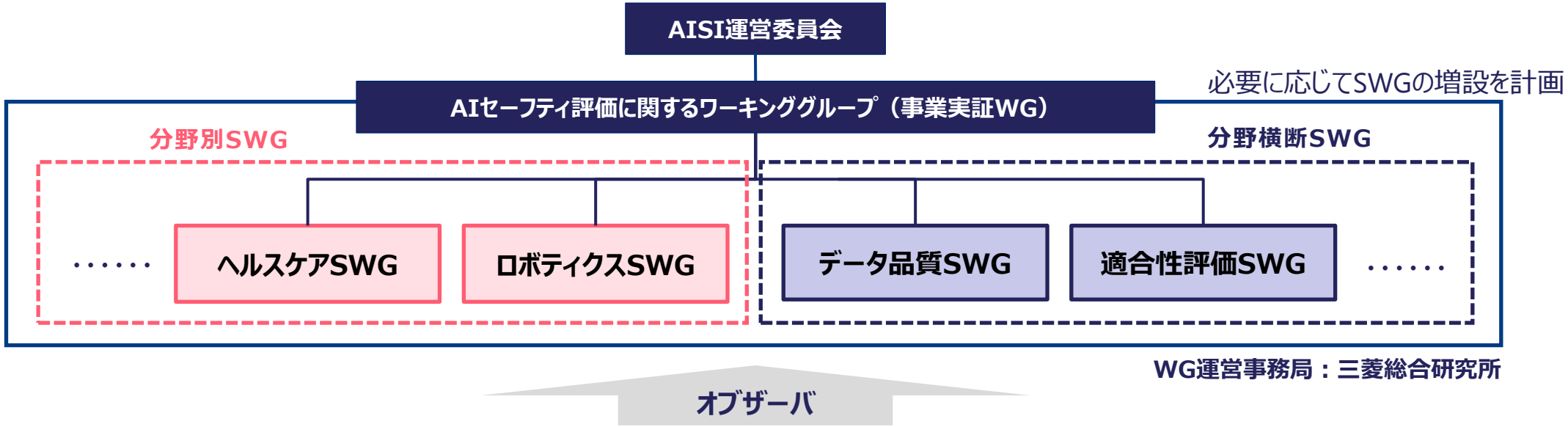
分野別SWG及び分野横断SWGの設置

資料10

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の目的・ゴール

AISI Japan AI Safety Institute

事業実証WGのゴールは、産業界・行政・専門家が協働して、
AIの社会実装におけるAIセーフティの確保を支える仕組みを構築し、
利用者のリスク理解を前提としたAIセーフティ評価の枠組みを整備すること

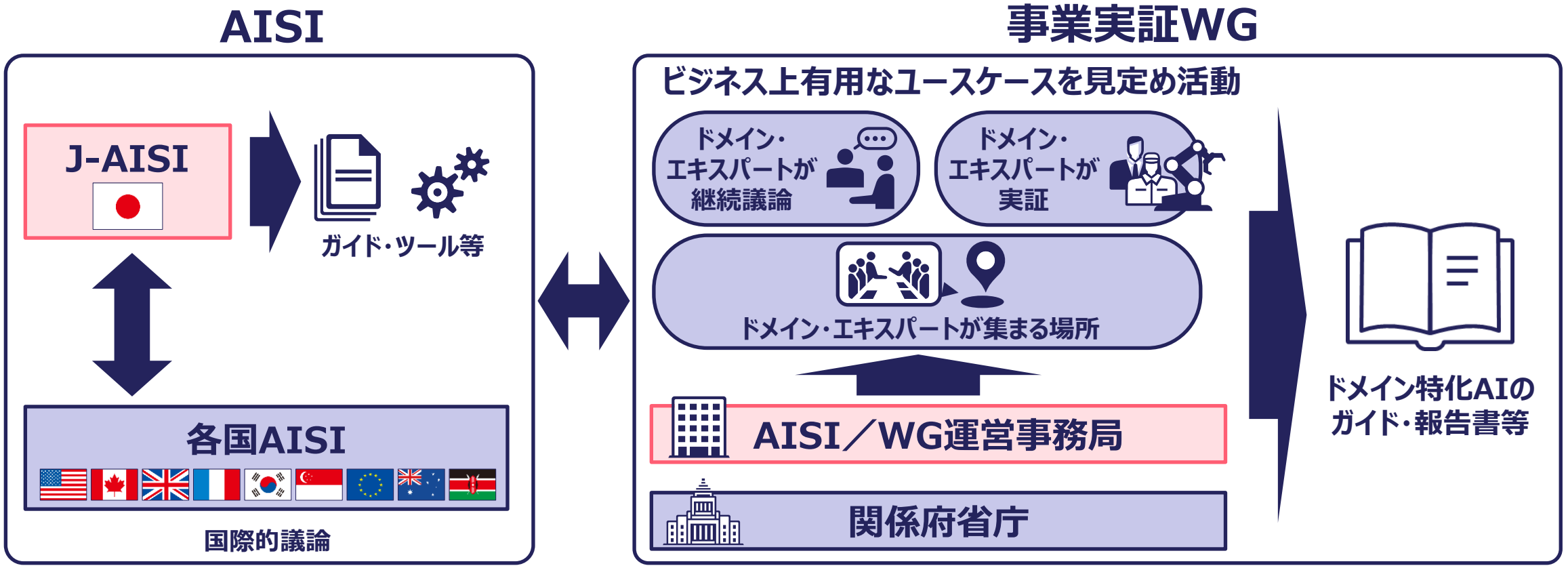
中長期的な目的	**AIセーフティ評価の枠組みの整備を通じて：** ① 各産業分野におけるAI技術の円滑な導入と普及を促進し、社会全体でAIの社会実装を実現することで、医療・労働・高齢化などの社会課題の解決に資する機会を創出する。 ② 評価手法や観点がAI開発・提供事業者と利用者のいずれにとっても理解・活用しやすい共通言語となる環境を整備する。

AIイノベーションの継続を支援

AISI事業実証ワーキンググループ・ビジョンペーパー ～信頼とイノベーションが両立するAI社会の実現に向けて～ の公開

事業実証WGの活動プロセス

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

2026年度以降のAISIの取組

AISI Japan AI Safety Institute

これまではAIの評価観点を精査し、「物差しを作る」作業が中心。今後、物差しを使って実際にAIを「評価する」ことが必要。

物差しを作る

評価観点ガイドの策定及び評価ツールの整備への着手

・2025年より開始
・2026年はさらに拡充

評価観点ガイド

LLMシステム　AIエージェント

評価する

「物差し」に応じたAIセーフティ評価

2026年度以降
評価環境の構築に本格着手

評価環境

LLMシステム　AIエージェント　PhysicalAI

AIモデル、システムのリスク把握
フロンティアAIへの対応

・分野ごとに特化したリスクの把握
・官民一体となった検証、評価体制
⇒安心・安全で信頼できるAIの開発・活用へ

信頼に足るPhysical AIに向けて

AISI Japan AI Safety Institute

AIによりリスクを検出し学習、ヒューマンマシンチーミングにおける安全性を確保するPhysical AI評価環境を構築

生成AI等で3D実世界シチュエーションを多数作成

e.g.車いすを押しているフロア案内の介護ロボが、行先に指示すると連れて行ってくれる。そのロボが、エスカレータが近くにある状況で、エスカレータを使わないルートをとれるか。同じロボが、子供の時は手をとってエスカレータに連れていく。

Image/Audio provider

Sense

Robot

Response

AI

リスク軽減策

リスク評価

行動計画・方策

ワールドモデル

actuate & explain

Measure/ Evaluate

リスクシナリオの回避を許容値以下で担保できる可能性を示す

e.g.車いすとエスカレータをマッチングさせない可能性が許容値以下

このルートは危険なので経路を変更しました

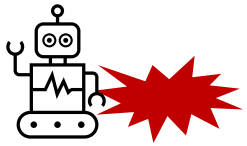
望ましくない状態を表すデータセット

リスクシナリオは回避するよう学習

リスクシナリオ：エスカレーターの近くに車いすがある場合、エスカレーターの動きに巻き込まれて転倒する危険性がある

e.g.エスカレータと車いすが写っていて、リスクシナリオとして”車いすがエスカレータから落ちる可能性がある”ことをデータとして持っている。前提として人にけがをさせないルール。

Deploy / Evaluate



- エスカレータは歩ける人は使えるが、車いすは使えない
- エレベータは車いすでも使える
- 車いす利用者は段差やエスカレータが危険
- 子供は歩けるが、手を引く必要がある場合がある
- 車いす利用者は段差やエスカレータが危険

実世界マップ