BRiDGE

研究開発とSociety 5.0との橋渡しプログラム

programs for **B**ridging the gap between **R**&d and the **ID**eal society (society 5.0) and **G**enerating **E**conomic and social value

令和７年度補正予算 研究開発等計画

AIセーフティ・インスティテュート（AISI）の抜本的機能強化の推進

令和８年１月
経済産業省

● 実施する重点課題（特に該当するものには◎、そのほかで該当するものがあれば○（複数可）を記載）

SIPや各省庁制度による研究開発成果の社会実装・市場開拓の加速化	他の戦略分野等との技術の融合による研究開発	スタートアップによるイノベーションの創出・促進	産学官を挙げた人材の育成・確保	グローバルな視点での連携強化
○	◎			

● 関連するSIP課題（該当するものに○を記載）

持続可能なフードチェーン	ヘルスケア	包摂的コミュニティ	学び方・働き方	海洋安全保障	スマートエネルギー	サーキュラーエコノミー	防災ネットワーク	インフラマネジメント	モビリティプラットフォーム	人協調型ロボティクス	バーチャルエコノミー	先進的量子技術基盤	マテリアル事業化・育成エコ

1．施策の概要

- 英国のAISIは、自らAI安全性・セキュリティに係る評価ツールを開発。また米国は、企業が作ったリスクの高いAIを評価・分析する能力も持つべく強力に体制強化。**我が国のAISIは、令和6年2月に設置されて以降、人員・予算が脆弱なため、AIの安全性・セキュリティに係る機能が十分でない**。このような中、世界では、生成AIに加えて、AIエージェントやフィジカルAIといった高性能なAIシステムの利用・開発が急速に進んできているが、我が国ではこうした最先端のAIシステムにおいて原因究明が困難なインシデント発生時に対応する体制が十分に整備されておらず、AI利用・開発の促進を阻害する可能性。
- 令和7年9月のAI法の全面施行を受けて策定された**AI基本計画でも、AIの利用・開発を支えるべく、AIモデルの適正性に係る評価機能の構築を含むAISIの抜本的な機能強化**が求められているほか、この度の**経済対策に係る総理大臣指示でもAI分野の官民連携投資が求められており、早急に民間企業等が安心して信頼できるAIを利用・開発できるよう機能の構築が不可欠。**
- そこで、我が国の民間企業等のAI利用・開発を促進しAIへの民間投資を増やすため、**AISIの抜本的な機能強化を前倒しで進め、以下の２つを実施する。**

（１）AIエージェントの利用・開発に係るガイドやツール等の開発

- 各国が参画するAISI国際ネットワーク会合を通じて、日本AISIが、AIエージェントをはじめとした最新のAIシステムの開発・利用におけるセキュリティ確保に係る課題等を把握。
- これらの結果をもとに、 AIエージェントをはじめとした最新のAIシステムが連動して動作するシステムにおけるセキュリティ確保のための調査（技術動向・事例・課題等）を実施。調査結果等に基づき、分野※ごとのAI開発・利用に係るガイドの作成 、及びガイドに基づくAIシステムに係る安全性検証を行うためのツールを開発。

※ガイド作成及びツール開発を行うドメインの想定例：分野共通、ヘルスケア分野、ロボット分野等のAIエージェントの活用が想定される業界

（２）AI評価に必要なベンチマークの開発

•信頼できるAIの開発に不可欠な、学習用（質問と答えのセット）と評価用（よく似た質問と答えのセット）それぞれのデータセット作成及び評価手法の確立※を行うとともに、妥当性検証のためのAI評価プログラムの開発等を行いベンチマークを開発。

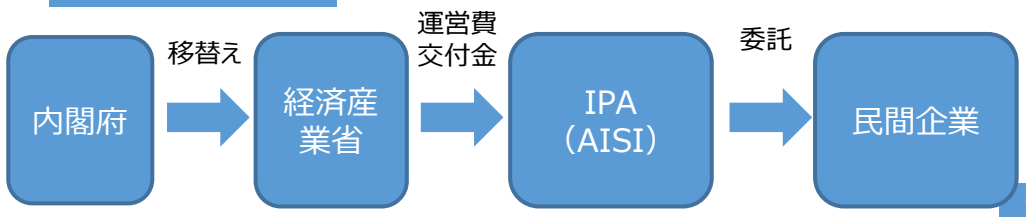
※ベンチマークを開発するカテゴリー（①安全性（バイアス、偽情報・誤情報、情報漏洩等）、②分野依存（ヘルスケア、法律、行政、経済等）、③Jailbreak、④セキュリティ、⑤エージェントモデル、⑥マルチモーダル、⑦ロボティクス、⑧評価プラットフォーム

- 国内外の類似の仕組みや認証スキームの検討に係る調査等を行い、これらの結果も参照しつつ、上記のベンチマーク等を用いてAIの安全性を認証するための枠組みを検討。

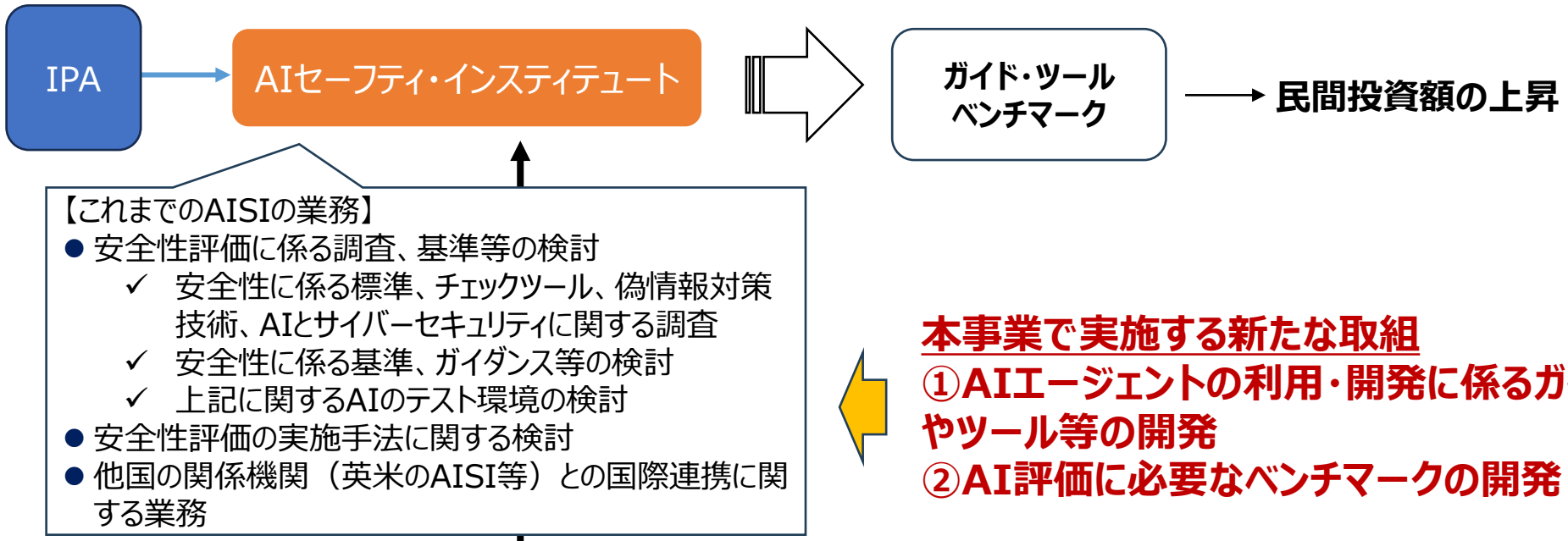
2．施策の対象・成果イメージ

生成AIのみならず、AIエージェント、フィジカルAIを含めた、**各ドメイン別のAIセーフティ・セキュリティのガイド・ツール・評価基盤（ベンチマーク、コンソーシアムによる検証制度）を構築**し、**AIへの民間投資額を上昇**させる。

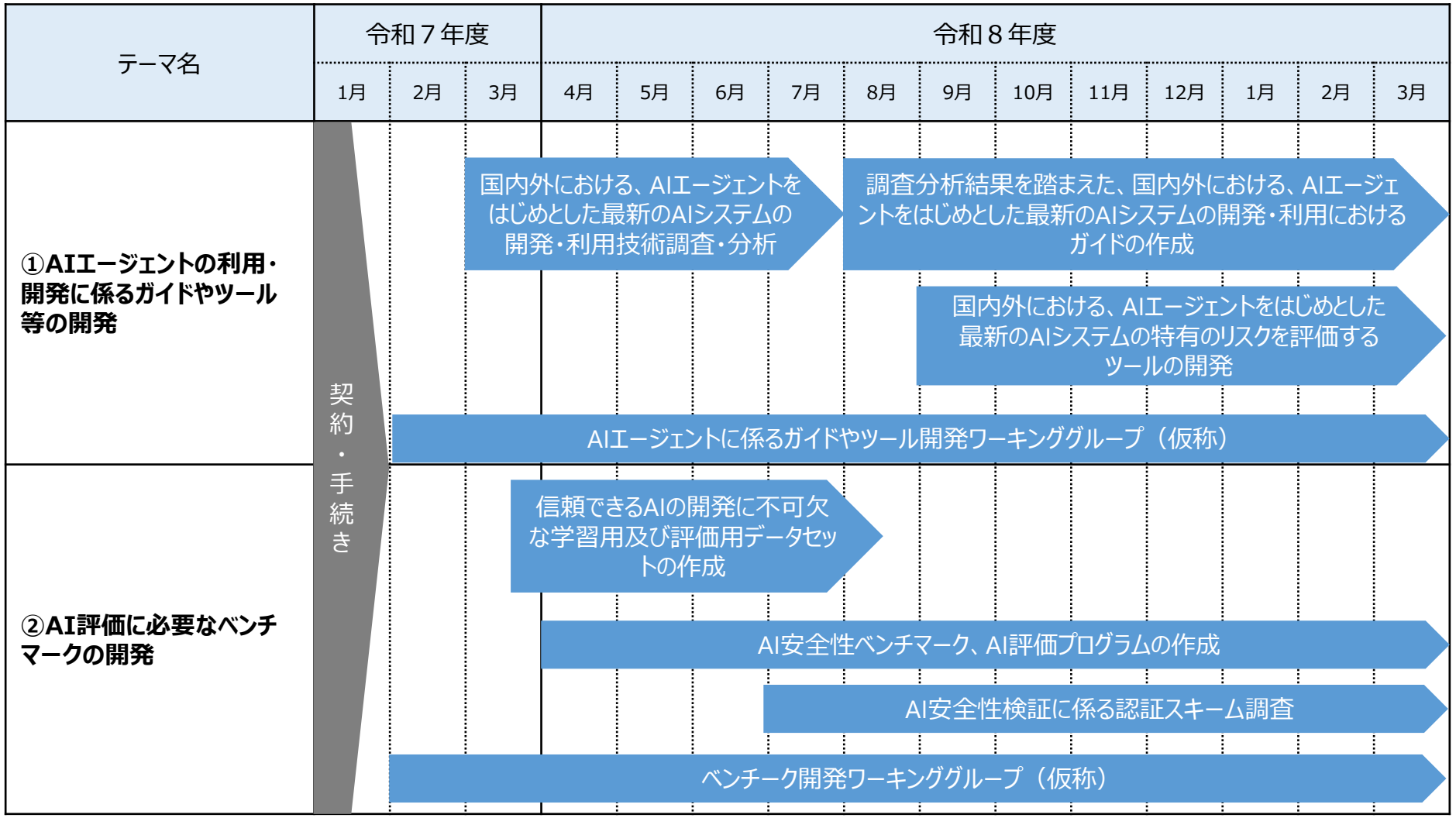
3．資金の流れ



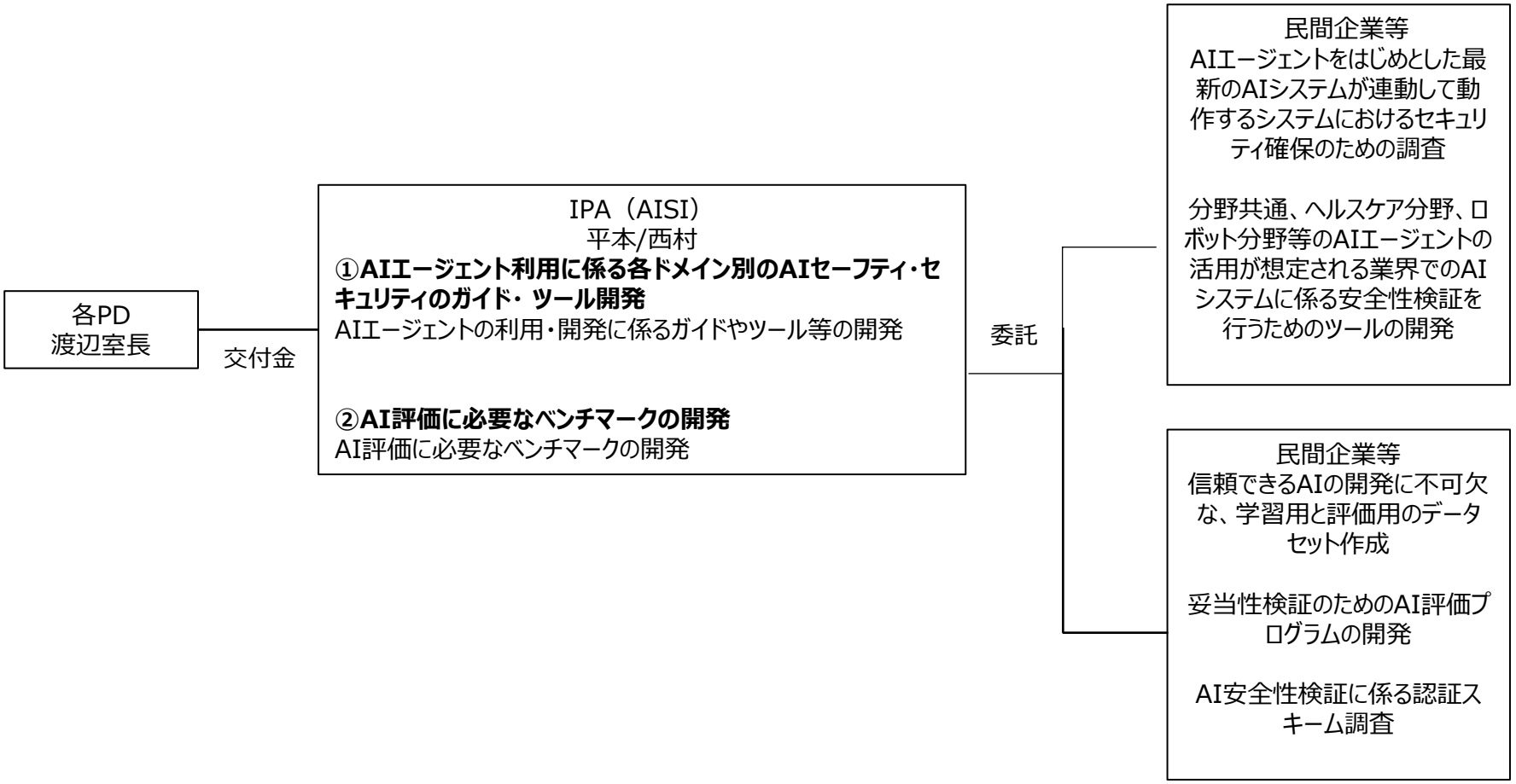
4．取組内容（システム概念図）



５．取組スケジュール


テーマ名
令和７年度
令和８年度
1月
2月
3月
4月
5月
6月
7月
8月
9月
10月
11月
12月
1月
2月
3月
①AIエージェントの利用・開発に係るガイドやツール等の開発
②AI評価に必要なベンチマークの開発
契約・手続き
国内外における、AIエージェントをはじめとした最新のAIシステムの開発・利用技術調査・分析
調査分析結果を踏まえた、国内外における、AIエージェントをはじめとした最新のAIシステムの開発・利用におけるガイドの作成
国内外における、AIエージェントをはじめとした最新のAIシステムの特有のリスクを評価するツールの開発
AIエージェントに係るガイドやツール開発ワーキンググループ（仮称）
信頼できるAIの開発に不可欠な学習用及び評価用データセットの作成
AI安全性ベンチマーク、AI評価プログラムの作成
AI安全性検証に係る認証スキーム調査
ベンチーク開発ワーキンググループ（仮称）

６．実施体制


各PD
渡辺室長
交付金
IPA（AISI）
平本/西村
①AIエージェント利用に係る各ドメイン別のAIセーフティ・セキュリティのガイド・ツール開発
AIエージェントの利用・開発に係るガイドやツール等の開発
②AI評価に必要なベンチマークの開発
AI評価に必要なベンチマークの開発
委託
民間企業等
AIエージェントをはじめとした最新のAIシステムが連動して動作するシステムにおけるセキュリティ確保のための調査
分野共通、ヘルスケア分野、ロボット分野等のAIエージェントの活用が想定される業界でのAIシステムに係る安全性検証を行うためのツールの開発
民間企業等
信頼できるAIの開発に不可欠な、学習用と評価用のデータセット作成
妥当性検証のためのAI評価プログラムの開発
AI安全性検証に係る認証スキーム調査

７．実施内容・到達目標（KPI）

テーマ名	実施内容の概要 到達目標（KPI）
①AIエージェントの利用・開発に係るガイドやツール等の開発	【実施内容】 • 国際AISIネットワークの構築と拡大の推進。AIエージェント等の最新のAIシステムに係る議論を行い必要な情報を把握 • 国内外における、AIエージェントをはじめとした最新のAIシステムの開発・利用技術調査・分析 • 調査分析結果を踏まえた、国内外における、AIエージェントをはじめとした最新のAIシステムの開発・利用におけるガイドの作成 • 国内外における、AIエージェントをはじめとした最新のAIシステムの特有のリスクを評価するツールの開発 【到達目標（KPI）】 • AIシステムの開発・利用におけるガイドの作成・提供　１件 • 最新のAIシステムの特有のリスクを評価するツールの開発・提供　１件
②AI評価に必要なベンチマークの開発	【実施内容】 • 信頼できるAIの開発に不可欠な学習用及び評価用データセットの作成 • 妥当性検証のためのAI評価プログラムの開発 • AI安全性検証に係る認証スキーム調査 【到達目標（KPI）】 • 信頼できるAIの開発に不可欠な学習用及び評価用データセットの作成・公開　３件 • 妥当性検証のためのAI評価プログラムの開発・公開　１件 • AI安全性検証に係る認証スキームの調査結果をもとにリスク分類や評価ガイドラインをもとにAIシステムを認証するためのスキームを構築