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６年度補正予算 研究開発等計画

AIセーフティ強化に関する研究開発構想

令和７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
											○		

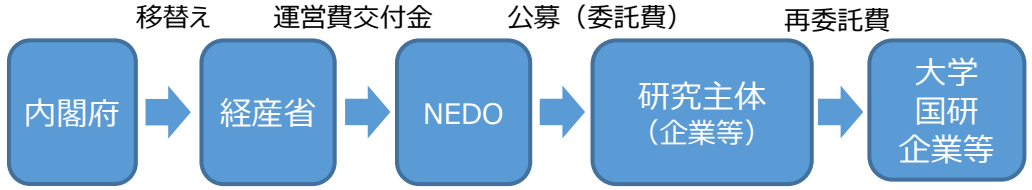
1．施策の概要

- 2023年11月の英国主催のAIセーフティ・サミット（わが国からは、岸田前総理がオンライン出席）を契機として、**各国はAIセーフティ・インスティテュート（AISI）を立ち上げ、AIセーフティの確保に関する議論を急速に進めている**。こうした国際的な情勢も踏まえ、わが国としても、**第7回AI戦略会議における総理指示を踏まえ、AIセーフティ・インスティテュート（AISI）を立ち上げ、国際的な議論に参画**している。本事業は、こうした**官民一体としての取組を研究開発の側面から支援**する。
- 特に、**安全であるかを見極める評価技術**は、生成AIを適切に管理・利用するために必要である。欧州AI法に代表されるように、**リスクベースのアプローチ**により、安全性を4段階のカテゴリで評価し対策を義務付けるといった法規制が進みつつある。2024年11月の米国（国務省、商務省）主催の国際AISIネットワーク会合につづき、2025年２月にはフランス主催のAIアクション・サミットが予定されており、今後AIセーフティについては**相当程度の高頻度で国際的な議論が進むと想定されることから、早急な技術開発を目指す**もの。
- リスクベースアプローチの基になる**安全性の“ものさし”（基準）となる技術（評価・管理技術）の開発（①AIセーフティ評価・管理基盤技術）**、**人間拡張技術などのサイバー空間とフィジカル空間をつなぐ領域での評価手法の開発と実証およびテスト環境構築技術の開発（②応用領域別AIセーフティ評価・実装技術）**、**国際標準化および普及のためのガイダンス等の整備（③AIセーフティ基準・ガイダンスと標準化）**を実施する。

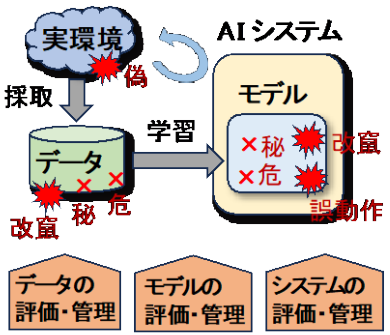
2．施策の対象・成果イメージ

- AIセーフティ基準のコアとなる安全性評価・管理基盤技術の確立。
- 人間拡張技術などサイバー空間とフィジカル空間をつなぐ「暮らし」領域での安全性評価技術・構築技術の確立。
- ISO/IEC等における規格化に向けた標準化活動の精力的な実施。

3．資金の流れ

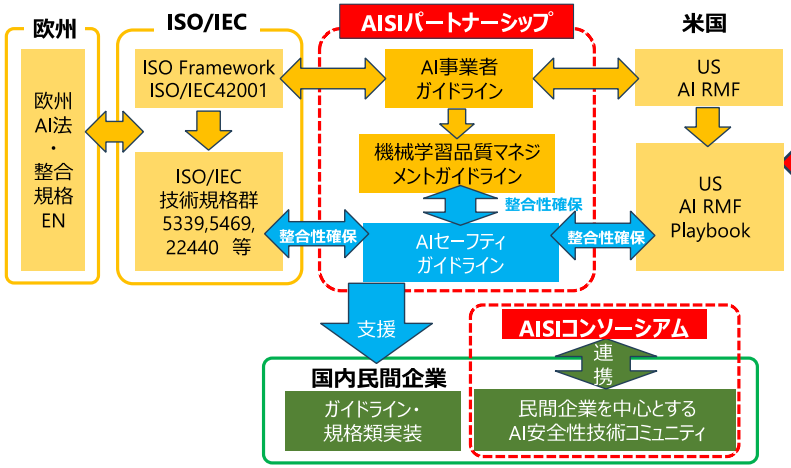


①AIセーフティ評価・管理基盤技術



AIモデル・システムに対する個別の攻撃や防御手法の研究は盛んだが、安全性の評価技術・管理技術が体系的に確立していない。AIセーフティ基準策定に必須。

③AIセーフティ基準・ガイダンスと標準化



・AIセーフティ基準の開発
・AIセーフティ基準の社会実装・普及の促進
・ISO/IECでの標準化活動と国際連携

②応用領域別AIセーフティの評価・実装技術

「暮らし」領域に特化した
- 評価シナリオやベンチマークデータセットを用いた安全性評価手法
- テスト環境構築技術
- 安全なシステムを構築するための基盤モデル技術

サイバー空間とフィジカル空間をつなぐ応用領域に特有のリスクに対応するためのAIセーフティ評価・実装技術を開発する。

・AIセーフティ基準のコアとなる安全性評価・管理基盤技術の確立
・人間拡張技術などサイバー空間とフィジカル空間をつなぐ領域での安全性評価技術・テスト環境構築技術の確立
・ISO/IEC等における規格化に向けた標準化活動の精力的な実施

研究開発項目① AIセーフティ評価・管理基盤技術

研究開発の概要と重要性

ChatGPTに代表される生成AI技術が、個人や社会に与えるリスクが高まっている。特に、実世界を扱うマルチモーダルAIは、生活空間で重要性を増しているが、人と関わるフィジカル空間での危険を増大させ、人とのインタラクションを通じて権利侵害や倫理違反を引き起こす恐れがある。生成AIのリスクを評価・管理するための基盤技術は、世界的にも研究開発途上にあるが、LLMやマルチモーダルAIを含む生成AIについてセーフティ基準を策定するために不可欠である。そこで、AIセーフティの評価・基盤技術の研究開発では、システムライフサイクル全体のリスクアセスメント技術を開発する目的で、データ、学習モデル、AIシステムのそれぞれの評価・管理技術を開発する。

期間内の達成目標

データ、学習モデル、AIシステムそれぞれのAIセーフティ評価・管理技術の開発にあたって、具体的には、評価観点と評価水準の整理およびこれらに対応する評価・管理手法の基礎設計、評価ツールのプロトタイプの試作およびその課題点分析、評価用ベンチマークデータセットが具備すべき要件整理、ベンチマークデータセットに基づく試験的安全性評価結果などの成果を目指す。また、AIセーフティ基準（体系化した評価観点・評価水準及び評価・管理手法）の検討・改定に役立てる目的で、上記の活動を整理した報告書を作成する。

研究開発の詳細（参考）

a)データの評価・管理技術

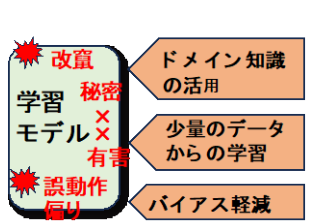
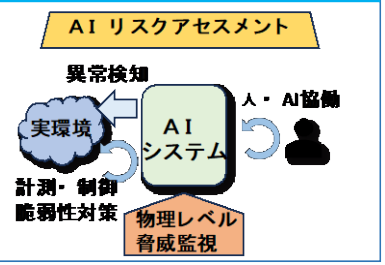
多様な環境や状況からのデータ収集する場合、時間経過に伴って環境や状況が変化する場合、多様なセンシング方法でデータ収集する場合、ドメイン知識や人間による解釈がデータに含まれない場合、データが不足している場合などについて、誤情報・危険情報・秘密情報などによるデータの汚染に対応できていない。こういった課題を解決するために、データ汚染の検知技術、構造化知識の活用によるデータセット構築技術、データ拡張技術、アラインメント技術について、評価観点及び評価水準の整理や評価・管理手法の研究開発を行う。

b)学習モデルの評価・管理技術

言語化されていないドメイン知識の不足による現実から乖離した出力（ハルシネーション）や、データ取得の環境や状況の知識の不足によるモデル動作の偏りについて、対応できていない。そこで、ドメイン知識の活用によるハルシネーションの軽減、汚染されていない少量データからの学習による誤動作の軽減、公平性の評価・実現のためのバイアスの軽減について、評価観点及び評価水準の整理や評価・管理手法の研究開発を行う。

c)システムレベルの評価・管理技術

AIシステムの誤動作を生じさせるような人や実環境からの異常入力・攻撃への対策や監視ができていない。また、人とAIシステムの間のインタラクションにおいて人とAIシステムの間で意図を共有できていないことから生じる誤動作への対策ができていない。そこで、AIシステムに対するシステムレベルのリスクを体系化し、評価観点及び評価水準を整理する。また、人や実環境からの異常入力や攻撃への異常検知・監視技術、人-AI協働における誤動作対策に関して、評価観点及び評価水準の整理や評価・管理手法の研究開発を行う。

研究開発項目	①-aデータの評価・管理技術	①-bモデルの評価・管理技術	①-cシステムレベルの評価・管理技術
			
技術的課題	データ汚染などの個別の攻撃手法の研究は盛んに行われているが、データの安全性評価手法・管理手法はまだ確立されていない。	誤動作や情報漏洩を生じさせる個別の攻撃手法・個別の対策技術の研究は盛んに行われているが、モデルの安全性評価手法・実現方法はまだ確立されていない。	攻撃手法の研究は盛んに行われているが、攻撃に晒されたAIシステムの安全性評価手法・防御手法はまだ確立されていない。

研究開発項目② 応用領域別AIセーフティ評価・実装技術

研究開発の重要性

生成AIがフィジカル空間で利用されるようになり、リスクが増大している。特に「暮らし」領域では、人と密接に関わるフィジカル空間での危険が増大し、人とのインタラクションにより、権利侵害・倫理違反が増加し、AIのリスクのインパクトが大きい。AIの利用方法、出力内容、安全性の判断指標は、AI を利活用する応用領域ごとに大きく異なり、応用領域ごとの具体的リスクに応じて、AI のセーフティの評価・実装技術を開発する必要がある。「暮らし」領域では、人の身体・精神・財産等への安全性やプライバシーなどへのリスクについて、AI セーフティの評価・実装技術を開発する必要がある。

研究開発の概要

そこで、「暮らし」領域に特有のAIセーフティを評価するための評価シナリオ（想定され得るリスク状況・動作のパターン）やベンチマークデータセットなどを用いた評価手法を開発し、実フィールドを模擬したテスト環境を構築する技術を開発し、安全なシステム開発のため基盤モデル技術を構築する。
実証する応用領域としては、サイバー空間とフィジカル空間での人間の行動や経験を拡張するAI技術（人間拡張AI技術）など、「暮らし」領域を対象とし、人間性に関わる評価観点に焦点を当てて取り組む。また、開発した安全性評価・実装技術についての知見を研究開発項目③のAIセーフティ基準へ反映させる。

期間内の達成目標

サイバー空間とフィジカル空間での人間の行動や経験を拡張するAI技術（人間拡張AI技術）などの安全性評価・実装技術を開発し、「暮らし」のユースケースを対象として、安全性評価に用いる実環境と仮想環境の構築技術を確立する。具体的成果として、事故データや不具合データを活用したAIセーフティ評価手法を構築し、「暮らし」領域の5件のAI製品・サービス等を対象として、評価手法の実証実験を行う。また、「暮らし」領域のAI技術のプライバシーの評価観点及び評価水準を策定し、多様なAI製品・サービス等への適用を通じて検証を行う。これらの評価手法、評価観点及び評価水準を活用し、1か所の現場においてAI技術導入の検証を実施する。

研究開発項目③ AIセーフティ基準・ガイダンスと標準化

研究開発の重要性

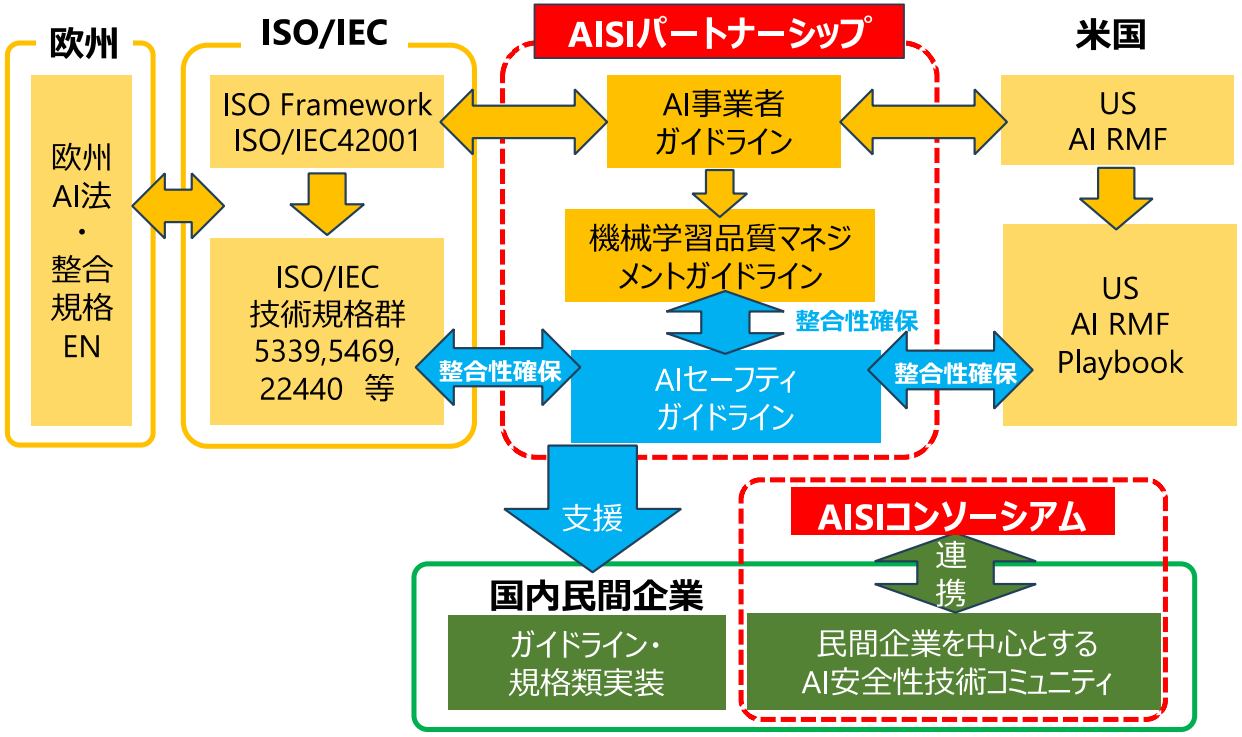
生成AI技術が個人や社会に与えるリスクに対応するために、AI技術の安全性評価が欧州AI法などによって義務付けられるようになり、AI技術の安全性評価に関して国際的なルール統一・整合の動きが盛んになっている。このため、日本のAI製品・サービスの普及にとって、日本の産業界が国際的なルール協調から取り残されないようにすることが不可欠である。具体的には、AI安全に関する国際標準を主導的に策定し、国際規格等と整合した形でAI開発現場に役立つAIセーフティガイドライン（AIセーフティ基準をまとめた文書）や実装解説等（ガイダンス）を作成し、広く普及させる必要がある。

研究開発の概要

AIセーフティガイドラインを策定し、企業向け実装支援施策・解説等を拡充し、ISO/IEC等での規格化、民間企業を中心とするAI安全性技術コミュニティの立ち上げによる社会実装・普及の促進を行う。また、研究開発項目①および研究開発項目②で得られた技術や知見を、AIセーフティ基準・ガイドラインに反映させる。

期間内の達成目標

AIセーフティガイドラインを公開・改訂する。また、AI安全性に関する企業コミュニティ支援を目的として、AISIコンソーシアムと連携し、企業向けの実装解説等を策定する。さらに、ISO/IECでのAI標準の国際的議論に貢献し、AISIに成果を提供する。



5．取組スケジュール

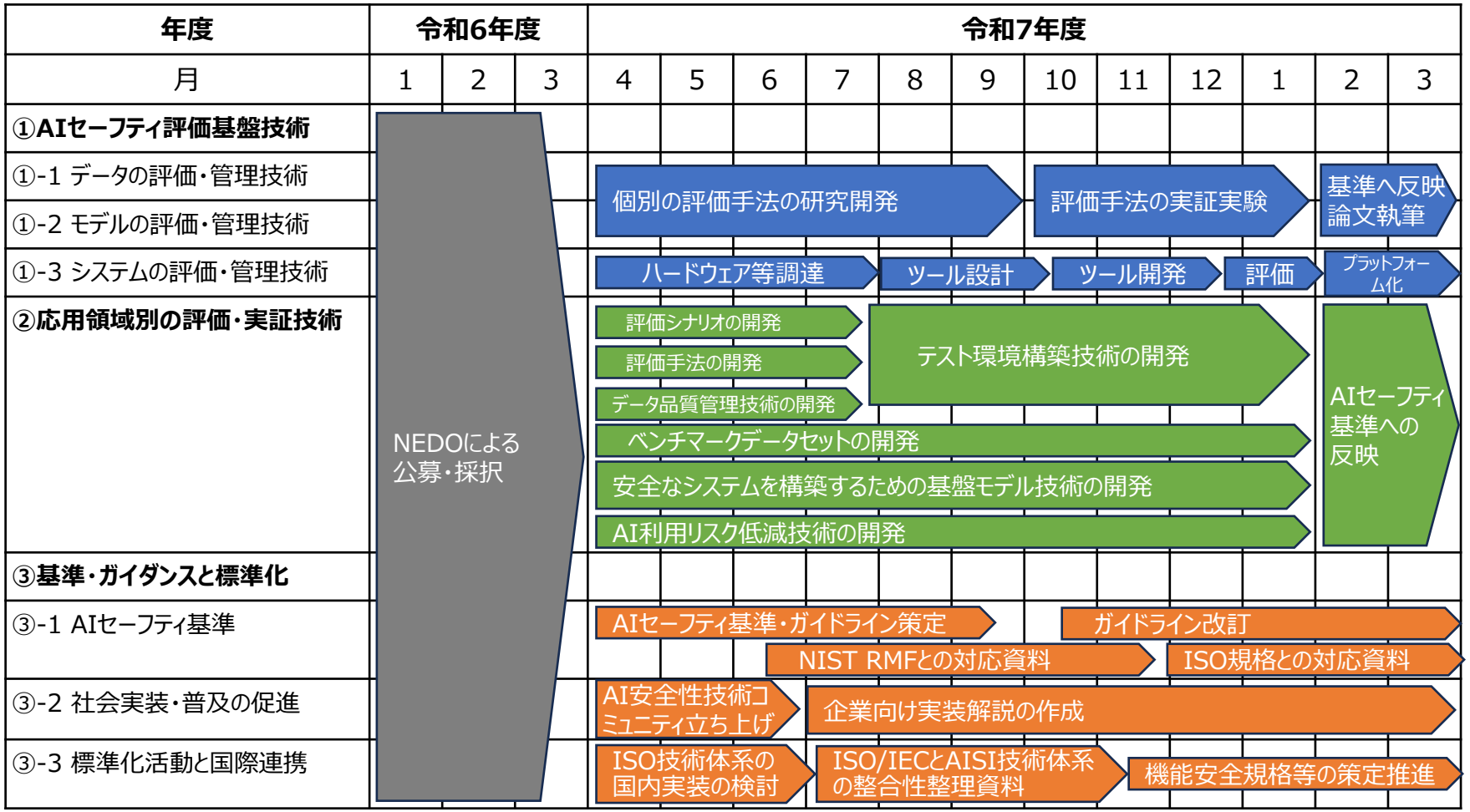
●成果

①AIセーフティ評価・管理基盤技術　AIセーフティ基準の詳細化を目的として安全性評価・管理の基盤技術を開発し、国際会議や国際誌で発表するとともに、AIセーフティ基準の策定に資する報告書を作成する。

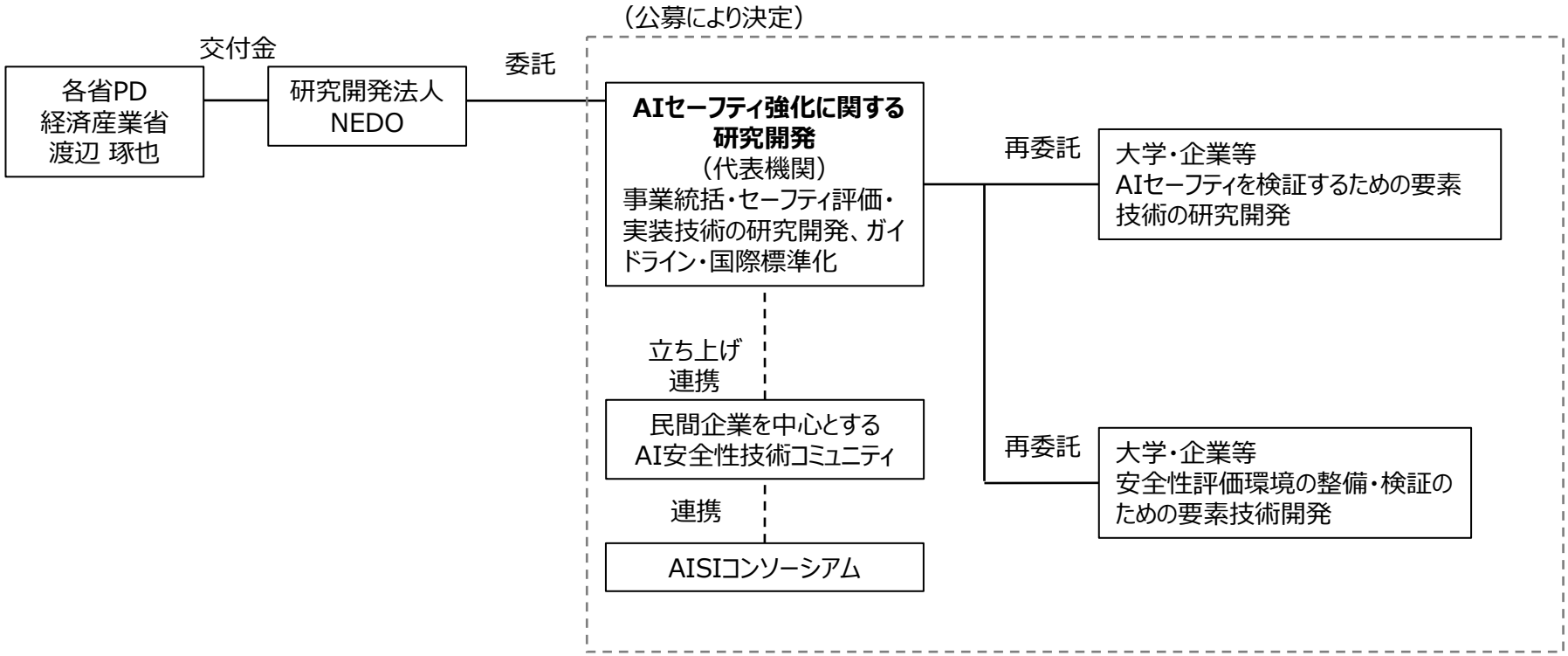
②応用領域別AIセーフティの評価・実装技術　「暮らし」領域のAI技術の評価シナリオやベンチマークデータセットを用いた安全性評価手法、テスト環境構築技術、安全なシステムを構築するための基盤モデル技術等を開発する。

③AIセーフティ基準・ガイダンスと標準化　AIセーフティガイドラインを策定し、民間企業を中心とするAI安全性技術コミュニティを通じて社会実装・普及促進のための企業向け実装支援施策・解説等を開発し、ISO/IEC等での標準化活動を行う。

●スケジュール



6．実施体制


（公募により決定）
交付金
委託
各省PD
経済産業省
渡辺 琢也
研究開発法人
NEDO
AIセーフティ強化に関する研究開発
（代表機関）
事業統括・セーフティ評価・実装技術の研究開発、ガイドライン・国際標準化
再委託
大学・企業等
AIセーフティを検証するための要素技術の研究開発
立ち上げ
連携
民間企業を中心とするAI安全性技術コミュニティ
再委託
大学・企業等
安全性評価環境の整備・検証のための要素技術開発
連携
AISIコンソーシアム

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セーフティ評価・管理基盤技術	【AIセーフティ基準の詳細化のコアとなる安全性評価・管理基盤技術の確立】 AIセーフティ基準の詳細化を目的として、AIセーフティの評価・管理基盤技術を確立する。これまでのAIセーフティ基盤技術は、個別の攻撃手法・防御手法の基礎研究段階にあったが、本事業により、基準作りに不可欠な体系的な安全性評価・管理基盤技術の構築段階に引き上げる。 事業終了時の具体的成果として、データ、モデル、システムの評価観点と評価水準の整理および評価・管理の要素技術の基礎設計、評価ツールのプロトタイプ試作およびその課題点分析、評価用ベンチマークデータセットが具備すべき要件整理、ベンチマークデータセットに基づく試験的安全性評価結果などが想定される。また、AIセーフティ基準の検討・改定に役立てる目的で、上記の活動を整理した報告書を作成する。
②応用領域別AIセーフティの評価・実装技術	【人間拡張など「暮らし」領域の安全性評価・実装技術・テスト環境構築技術の確立】 サイバー空間とフィジカル空間での人間の行動や経験を拡張するAI技術等の安全性評価・実装技術を開発し、「暮らし」のユースケースを対象として、安全性評価に用いる実環境と仮想環境の構築技術を確立する。 事業終了時の具体的成果として、事故データや不具合データを活用したAIセーフティ評価手法を構築し、「暮らし」領域の5件のAI製品・サービス等を対象として、評価手法の実証実験を行う。また、「暮らし」領域のAI技術のプライバシーの評価観点及び評価水準を策定し、多様なAI製品・サービス等への適用を通じて検証を行う。これらの評価手法、評価観点及び評価水準を活用し、1か所以上の現場においてAI技術導入の検証を実施する。
③AIセーフティ基準・ガイダンスと標準化	【AIセーフティ基準の策定と普及・ISO/IECに向けた標準化活動の実施】 2024年夏以降、国際AISIネットワーク等においてAIセーフティに係るルール形成の動きが加速しているが、日本が国際的なルール協調から取り残されると、日本のAI製品・サービスの普及の障害となる。そこで、本事業により、AIセーフティ基準の策定や社会実装・普及の促進、ISO/IEC等での標準化活動を行い、国際AISIネットワーク等を通じて、生成AI向けAIセーフティ標準化のための国際協調を加速させる。 事業終了時の具体的成果として、AIセーフティ基準をまとめたAIセーフティガイドラインを公開・改訂する。また、AI安全性に関する企業コミュニティ支援を目的として、AISIコンソーシアム（事業実証WG）と連携し、企業向けの実装解説等を作成する。さらに、ISO/IEC等でのAI標準の国際的議論に貢献し、AISIに成果を提供する。