AIによるイノベーション実現に向けた AISIの取り組み

2025-03-15

AIセーフティ・インスティテュート（AISI：エイシー）
平本健二

AISI Japan AI Safety Institute

統合イノベーション戦略における３つの強化方策

AISI Japan AI Safety Institute

（１）重要技術に関する統合的な戦略

①コア技術の開発、他の戦略分野との技術の融合による研究開発（産学官の連携、AI・ロボティクス・IoT等による研究開発推進等）

②国内産業基盤の確立、スタートアップ等によるイノベーション促進（ユースケースの早期創出、拠点・ハブ機能の強化等）

③産学官を挙げた人材の育成・確保（産業化を担う人材、市場開拓を担う人材、研究開発を担う人材の育成・確保等）

（２）グローバルな視点での連携強化

①重要技術等に関する国際的なルールメイキングの主導・参画（開発・利用の促進、安全性確保、プレゼンスの確保等）

②科学技術・イノベーション政策と経済安全保障政策との連携強化（国際協力・国際連携を含めた戦略的な研究開発、技術流出防止等）

③グローバルな視点でのリソースの積極活用、戦略的な協働（国際頭脳循環の拠点形成、国際科学トップサークルへの参画等）

（３）AI分野の競争力強化と安全・安心の確保

①AIのイノベーションとAIによるイノベーションの加速（研究開発力の強化、AI利活用の推進、インフラの高度化等）

②AIの安全・安心の確保（ガバナンス、安全性の検討、偽・誤情報への対策、知財等）

③国際的な連携・協調の推進（広島AIプロセスの成果を踏まえた国際連携等）

（３）ＡＩ分野の競争力強化と安全・安心の確保

AISI Japan AI Safety Institute

- 生成AIはインターネットにも匹敵する技術革新とされ、社会経済システムに大きな変革をもたらす一方で、偽・誤情報の流布や犯罪の巧妙化など様々なリスクも指摘され、安全・安心の確保が求められる。
- 米国企業等の高性能・大規模な汎用基盤モデルが先行する中、我が国もそれに追随すべく計算資源の整備や大規模モデルの開発が進んでおり、また、小規模・高性能なモデルや複数モデルの組合せの開発など、新たな研究も進んでいる。
- AIはあらゆる分野で利用され、AIの開発や利活用等のイノベーションが社会課題の解決や我が国の競争力に直結する可能性がある。我が国においては、生成AIを含むAIの様々なリスクを抑え、安全・安心な環境を確保しつつ、イノベーションを加速する好循環の形成を図っていく。加えて、我が国が主導する広島AIプロセス等を通じて、今後も国際的にリーダーシップを発揮していく。

① ＡＩのイノベーションとＡＩによるイノベーションの加速
- 研究開発力の強化（データ整備含む）
- ＡＩ利活用の推進
- インフラの高度化
- 人材の育成・確保

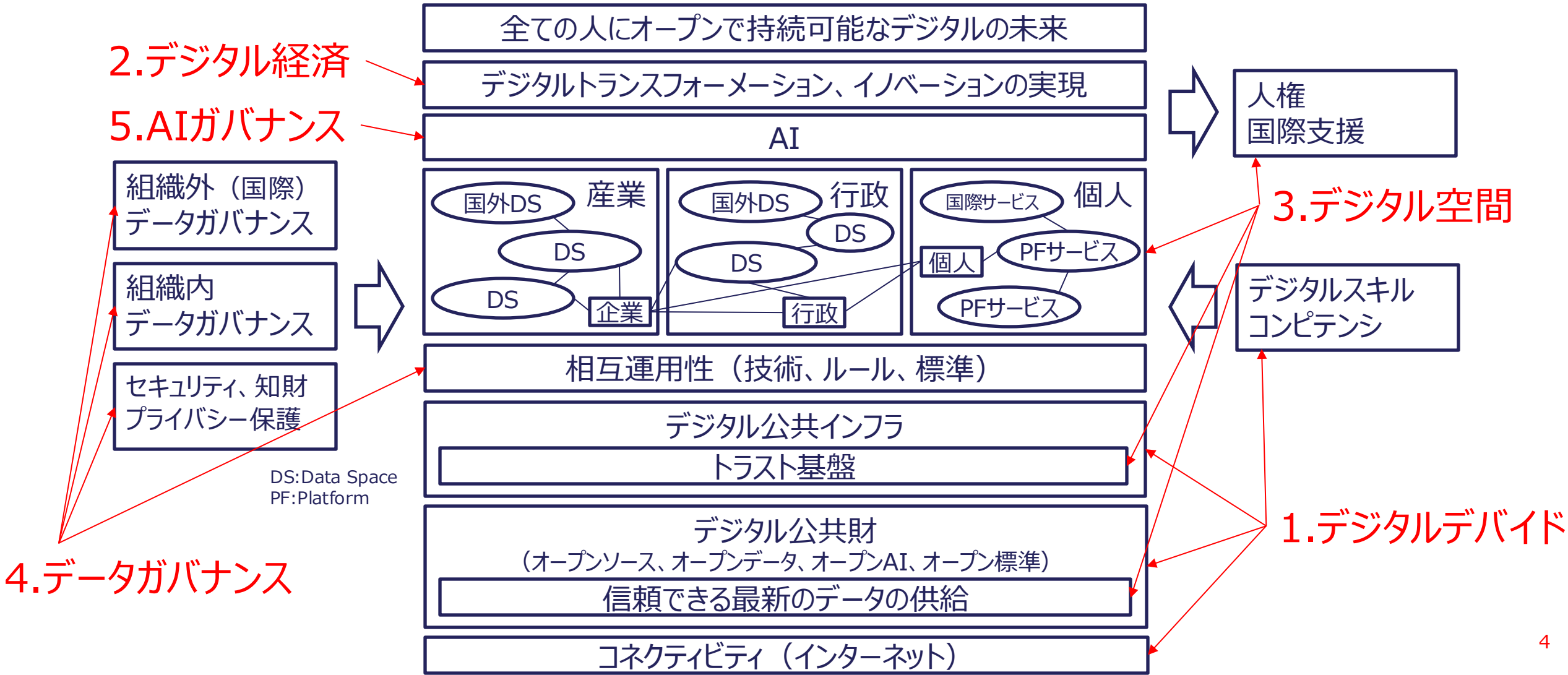
② ＡＩの安全・安心の確保
- 自発的ガバナンスと制度の検討
- ＡＩの安全性の検討
- 偽・誤情報への対策
- 知的財産権等

③ 国際的な連携・協調の推進

国連のGDCと同じゴールを目指している

※GDC：Global Digital Compact（2024-09-22）

- 2030年に、このゴールに到着しているのはデジタル社会の必要条件



AISIとは（AIセーフティ・インスティテュート）

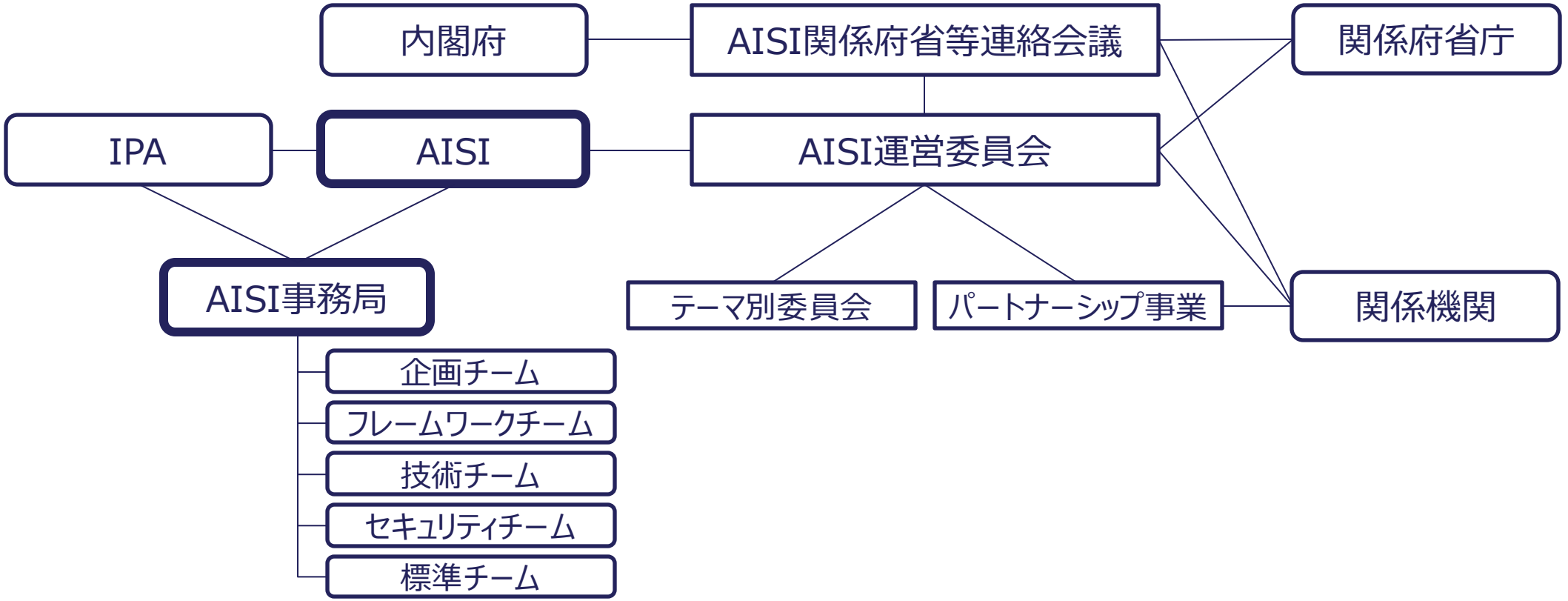
AISI Japan AI Safety Institute

- AISIは、内閣府を中心に10府省、5政府関連機関が連携する**官民の取組を支援する機関**である。（2024年2月設立。独立行政法人情報処理推進機構（IPA）に事務局）

- 役割
  - 政府への支援として、AIセーフティに関する調査、評価手法の検討や基準の作成等の支援を行う
  - 日本におけるAIセーフティのハブとして、産学における関連取組の最新情報を集約し、関係企業・団体間の連携を促進する。
  - さらに、 他国のAIセーフティ関係機関と連携する。
- 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

AISIの推進体制

- 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。
- AISIの中に、AISI所長を委員長とする「AISI運営委員会」を設置。
  - 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」（研究機関等との連携）を設置。



【関係府省庁】
・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】
・情報処理推進機構（IPA）
・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

LLMの安全性確保に向けた取り組み

AISI Japan AI Safety Institute

AI事業者ガイドライン
（総務省、経済産業省(2024-04)）

Cross Walk project
（米国AI Risk Management Framework との相互確認）

AIセーフティに関する評価観点ガイド
（AISI(2024-09)）

1. はじめに
2. AIセーフティ
3. 評価観点の詳細
4. 評価実施者及び評価実施時期
5. 評価手法の概要
6. 評価に際しての留意事項

AIセーフティにおける重要要素	AIセーフティ評価の観点									
	有害情報の出力制御	偽誤情報の出力・誘導の防止	公平性と包摂性	ハイリスク利用・目的外利用への対処	プライバシー保護	セキュリティ確保	説明可能性	ロバスト性	データ品質	検証可能性
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						●				
透明性		●	●				●	●	●	●

AIセーフティに関するレッドチーミング手法ガイド
（AISI(2024-09)）

1. はじめに
2. レッドチーミングについて
3. LLMシステムへの代表的な攻撃手法
4. 実施体制と役割
5. 実施時期及び実施工程
6. 実施計画の策定と実施準備
7. 攻撃計画・実施
8. 結果のとりまとめと改善計画の策定

レッドチーミングの種類

➢ レッドチーミングは以下のように分類できる。

攻撃計画・実施者が保有する前提知識の有無・程度による分類

- **ブラックボックステスト**
  （内部構造等の情報を未知としてレッドチーミングを行う）
- **ホワイトボックステスト**
  （内部構造等の情報を既知としてレッドチーミングを行う）
- **グレーボックステスト**
  （内部構造等の情報を一部既知としてレッドチーミングを行う）

レッドチーミングを実施する環境による分類

- **実運用環境**
  （AIシステムが実際に実用に供される運用環境）
- **ステージング環境**
  （実運用環境とほぼ同様の状態でテストや不具合のチェック等を行う環境）
- **開発環境**
  （AIシステムの開発を行う環境）

レッドチーミング実施において攻撃シグネチャを試行する方法による分類

- **自動化ツールによるレッドチーミング**
- **手動によるレッドチーミング**
- **AIエージェントを用いたレッドチーミング**

LLMシステムへの代表的な攻撃手法

➢ LLMシステムへの代表的な攻撃手法例として、下記が存在する。
これらを念頭に置いてレッドチーミングを行うのが望ましい。

- **直接プロンプトインジェクション**
  攻撃者が、悪意あるプロンプトをAIシステムに直接注入する攻撃
- **間接プロンプトインジェクション**
  攻撃者が、悪意あるプロンプトをAIシステムに間接的に注入する攻撃
- **プロンプトリーキング**
  攻撃者が、設定されたシステムプロンプトを引き出す攻撃
- **ポイズニング攻撃**
  攻撃者が細工したデータ・モデルを、訓練時に利用するデータ・モデルに紛れ込ませる攻撃
- **回避攻撃**
  AIシステムへの入力に悪意ある変更を加え、意図していない動作を引き起こす攻撃
- **モデル抽出攻撃**
  入出力の分析により、対象システムのモデルと同等の性能を持つモデルを作成する攻撃
- **メンバーシップ推論攻撃**
  入出力の分析により、あるデータが訓練データに含まれるかを特定する攻撃
- **モデルインバージョン攻撃**
  入出力の分析により、訓練データに含まれる情報を復元する攻撃

セキュリティ、データ品質、標準化等への取り組み

AISI Japan AI Safety Institute

- NIST AIリスク・マネジメント・フレームワーク及びプレイブックの日本語訳
- 多言語、多文化環境におけるAIの評価
- セキュリティ調査（AI利用時のセキュリティ脅威・リスク調査報告書）
- AI政策、活用動向調査　（DX白書、IPAwebサイト等）
- 「DX推進スキル標準」に生成AIに関する補記などを追加
- AIセーフティに関する活動マップとターミノロジー
- AIシステムのためのデータ品質管理ガイドブック（ドラフト）
- AI認証のためのJoint Certificationの検討
- 政府による適切なAIの調達・利用の在り方等に関する検討
- 国際活動、国内普及活動

グローバルな展開

AISI Japan AI Safety Institute

- AIやAIで活用するデータは、グローバルに展開されるため、国際的な協調やルールメイキングが重要となる。

各国AISI及び類似機関の連携

グローバルなアカデミック、プライベート・セクターとの連携

グローバルな情報収集、発信

AISI

Japan AI Safety Institute

AISIは、世界トップレベルのフィールドでの活躍を目指す人材を募集しています。