AIST
Create the Future, Collaborate Together

（資料1）

AIセーフティーと リスクマネジメントについて

2025年　11月19日

妹尾　義樹
標準化オフィサー
知財標準化推進部/インテリジェントプラットフォーム研究部門
国立研究開発法人　産業技術総合研究所

妹尾 義樹　産業技術総合研究所　知財標準化推進部　標準化オフィサー

経歴：　1986年 京都大学工学研究科　情報工学専攻修了（博士(工学)1996年）
2018年 産業技術総合研究所、人工知能研究企画室長
2020年 同所デジタルアーキテクチャ推進センター、情報標準化推進室長
2022年 同所知財標準化推進部　標準化オフィサー
2024年　AI品質マネジメントイニシアティブ会長

主な業務経歴：

2011年から2017年まではAI技術を活用したAnalyticsビジネス開発（主に北米市場向け）を企業にて指揮。2018年産総研入所後は人工知能研究企画、デジタル化に関わる標準化活動を指揮する傍ら、機械学習品質マネジメントについてのNEDOプロジェクトを研究代表として立ち上げる。また、標準化オフィサーとして産総研のIT関連標準化活動を統括。

AI品質マネジメントへの興味：

AIのビジネス活用にリスクテイクは必須。リスクテイクのための「ここまでやればOK」の基準作りをやりたい。日本が「品質」の強みでAIビジネスの主導権を取れるように。

内容

■AIの制御に関する世界の動向
■AIの品質マネジメントの課題
■生成AIの登場とAIセーフティーへの取り組み
■生成AIの品質マネジメントについて
■生成AIの発展と消費者にかかわる課題

なぜ AI に品質が必要か？

■ソフトウェアによる複雑な制御に
人命を預ける時代

- 航空機・鉄道車両
- 自動車
- 消費者機械
- インフラ（電力等）
- 医療・ヘルスケア

×

■ソフトウェア構築における機械学習AIへの依存度が高まっている

社会からみた AI への恐怖と要求

- 2019年頃から: AIに対する社会からの要求の明文化の動き

- **人間中心のAI社会原則**
  (2019. 3　統合イノベーション戦略推進会議)
  - 人間中心の原則
  - 教育・リテラシーの原則
  - プライバシー確保の原則
  - **セキュリティ確保の原則**
  - 公正競争確保の原則
  - **公平性**・説明責任及び透明性の原則
  - イノベーションの原則

- **OECD Principles on AI**
  (2019. 5. 22)
  - 全ての人への普遍的利益
  - **公平性と公正性の確保**
  - 透明性の確保と責任ある開示
  - **堅牢・セキュア・安全性**と
    リスクアセスメント
  - 開発運用者の責任

社会からみた AI への恐怖と要求

■2020年代: 法律・ガイド層の取り組み

米国

NIST
AIリスクマネジメント
フレームワーク
(2021.7作成開始)

米国政府調達要件に入ると
サプライチェーンに連なる
日本企業にも影響する恐れ

欧州

欧州委員会
2021.4.21法案公表
2024.8 **発効**

高リスクAI応用を特定
施行されると欧州市場向け
ビジネスで必須に

日本

(2021.7.9) 経産省
AIガバナンス・ガイドライン
(2024.4.19)
AI 事業者ガイドライン
▼
(2025.2) AI基本法案
技術革新とリスク対応のバランスを狙う

欧州AI法案

■ 欧州域内提供・**欧州域内の市民対象**のAIに広く適用

■ リスクベースでAIを4分類しそれぞれに規制（下記）

■ 汎用目的型AIモデル：モデル構築の計算量やユーザ数が閾値を超えたものについて規制（2025.8.2施行）

2025.2.2

Unacceptable RIsk

サブリミナル
公的なソーシャル
スコアリング等

2026.8.2

High Risk

- **リスク管理プロセス**
- **文書化・透明性確保**
- **人間による監視**
- **整合規格・共通仕様**
- **適合性評価手続**
- **データベース登録義務**
- **適合宣言・CEマーキング**
- **市場投入後のモニタリング**

などなど

Limited Risk

- 透明性確保
- 人に対する通知
- 影響の利用者への警告
（Deep Fake 等）

Minimal Risk

自主的な
取り組みを推奨

医療など規制産業はすべてここ

AI事業者ガイドライン

具体論の大半が
産総研のAI品質マネジメントガイドラインを引用

内閣府
人間中心のAI社会原則
「AI-Readyな社会」において、我が国社会全体、さらには多国間の枠組みで実現されるべき社会的枠組みに関して7つの原則にまとめたもの

総務省
AI利活用ガイドライン
AIの便益の増進及びリスクの抑制のため留意すべき10件の「AI利活用原則」の解説をまとめたもの

経済産業省
AI原則実践のためのガバナンスガイドライン
「アジャイル・ガバナンス・フレームワーク」に沿って、ガバナンスモデルを回していくための「行動目標(23件)」とその実践例

AI事業者ガイドライン第1.0版の構成 (24/4/19)

本編 **理念(Why)/原則(What)**

第1部 AIとは
第2部 AIにより目指すべき社会と各主体が取り組む事項
- A) 基本理念
- B) 原則
- C) 共通の指針
- D) 高度なAIシステムに関する事業者に共通の指針
- E) ガバナンスの構築

第3部 AI開発者に関する事項
第4部 AI提供者に関する事項
第5部 AI利用者に関する事項

別添 **具体論(What/How)**

第1部関連
- A) AIに関する前提
- B) AIによる便益/リスク

第2部関連
- A) 経営層によるAIガバナンスの構築
- B) AIガバナンスの事業者取組事例

第3部～第5部関連
- ・対応する本編の解説
- ・「共通の指針」の解説
- ・(第3部のみ)高度なAIシステム開発にあたって遵守すべき事項

契約ガイドラインの留意事項
チェックリスト
主体横断的な仮想事例
海外ガイドライン等の参照先

画像出典　https://www.nttdata.com/jp/ja/trends/data-insight/2025/050702/

AI事業者ガイドラインの利用者に関する記述

本編 別添
第5部

AI利用者に関する事項

- AI利用者は、AI提供者が意図した範囲内で継続的に適正利用、必要に応じたAIシステムの運用を行うことが重要であり、より効果的なAI利用のために必要な知見を習得することが期待されます

AIシステム・サービス利用時	U-2) i.	安全を考慮した適正利用	- AI提供者が定めた利用上の留意点を遵守して、**AI提供者が設計において想定した範囲内で利用**する - AIの出力について精度及びリスクの程度を理解し、**様々なリスク要因を確認した上で利用**する
	U-3) i.	入力データ又はプロンプトに含まれるバイアスへの配慮	- 公平性が担保されたデータの入力を行い、プロンプトに含まれるバイアスに留意して、責任をもって**AI出力結果の事業利用判断を行う**
	U-4) i.	個人情報の不適切入力及びプライバシー侵害への対策	- AIシステム・サービスへ個人情報を不適切に入力しないよう注意を払う - AIシステム・サービスにおける**プライバシー侵害に関して適宜情報収集**し、防止を検討する
	U-5) i.	セキュリティ対策の実施	- AI提供者による**セキュリティ上の留意点を遵守**する - AIシステム・サービスに機密情報等を不適切に入力しないよう注意を払う
	U-6) i.	関連するステークホルダーへの情報提供	- 公平性が担保されたデータの入力を行い、プロンプトに含まれるバイアスに留意して、**出力結果を取得し、結果を事業判断に活用した際は、その結果を関連するステークホルダーに合理的な範囲で情報提供**する
	U-7) i.	関連するステークホルダーへの説明	- AIの特性や用途、データの提供元となる関連するステークホルダーとの接点、プライバシーポリシー等を踏まえ、データ提供の手段、形式等について、あらかじめ**当該ステークホルダーに平易かつアクセスしやすい方法で情報提供**する - AIの出力結果を特定の個人又は集団に対する評価の参考とする場合は、人間による合理的な判断のもと、説明責任を果たす - **関連するステークホルダーからの問合せに対応する窓口を合理的な範囲で設置**し、AI提供者とも連携の上説明及び要望の受付を行う
	U-7) ii.	提供された文書の活用と規約の遵守	- AI提供者から提供されたAIシステム・サービスについての**文書を保管・活用**する - AI提供者が定めた**サービス規約を遵守**する

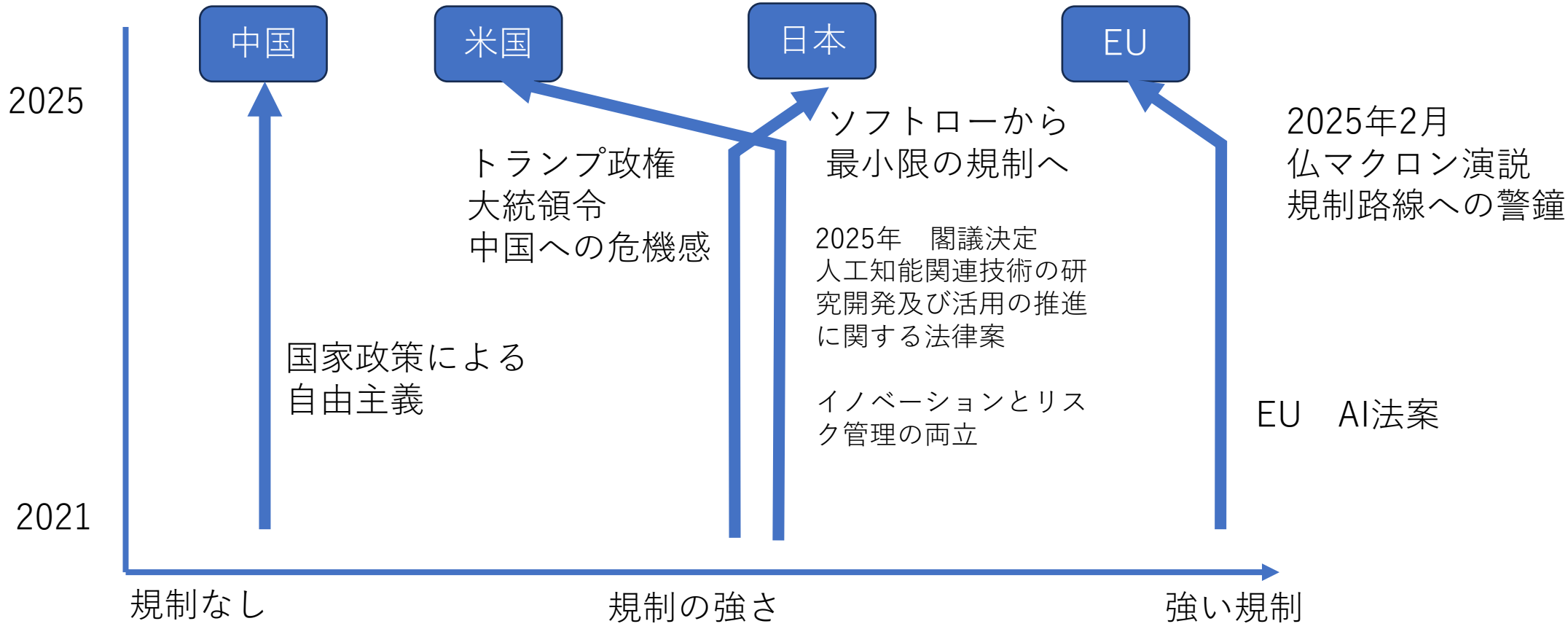
高度な AI システムを取り扱うAI利用者は、「第２部D. 高度なAIシステムに関係する事業者に共通の指針」について、I）～XI）を適切な範囲で遵守し、XII)について遵守すべき

24

AI規制についての最新の動向

■EU、中国、米国、日本のそれぞれの立ち位置と変化

- AIを国家成長戦略の重要技術との見立て

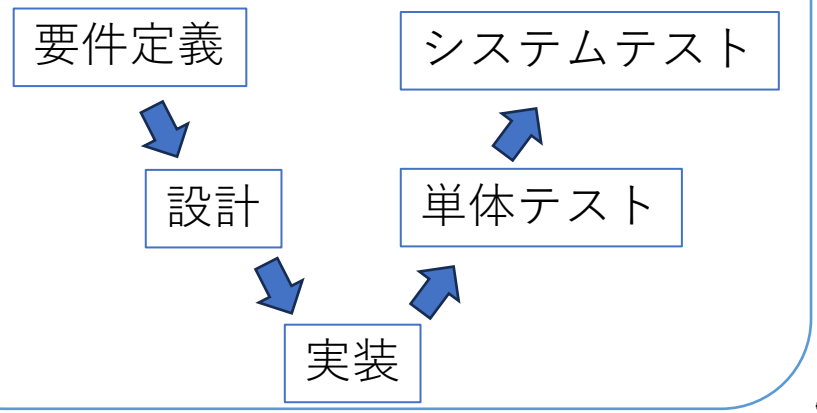


機械学習ＡＩの品質にまつわる課題

データ（お勉強の材料）が機能を決めるAIの品質管理をどうやるか？

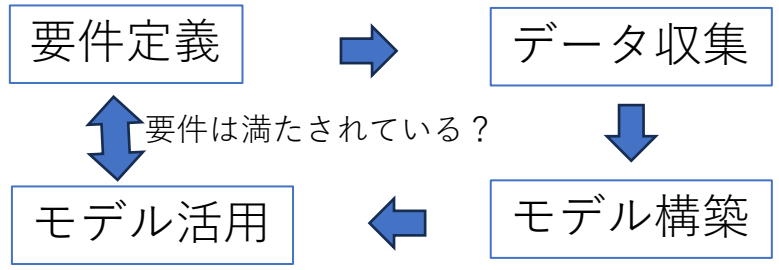
従来のソフトウェア品質保証
どうやればいいかを作りこみ確認

分析的アプローチ



機械学習AI
「お勉強」による機能の作りこみ

帰納的アプローチ



企業側のジレンマ：「一定の品質」を担保・説明できない

AIST
Create the Future, Collaborate Together

AI品質の難しさ：“猫問題”

猫の画像、を定義するのは難しい
- 一般的猫の特徴（髭、目、耳、尻尾…）
- 猫の種類（三毛猫、ペルシャ猫、マンチカン…）
- 光の当たり方、撮影角度、猫の姿勢…
- 漫画、顔だけ、しっぽだけ

何に使うか（＝要件）の定義が重要！

■機械学習を使えば、
- 検出ソフトウェアを容易に作成可能
- 訓練画像を集めて学習させるだけ！
  ⇒**判断基準は人には分からないまま**

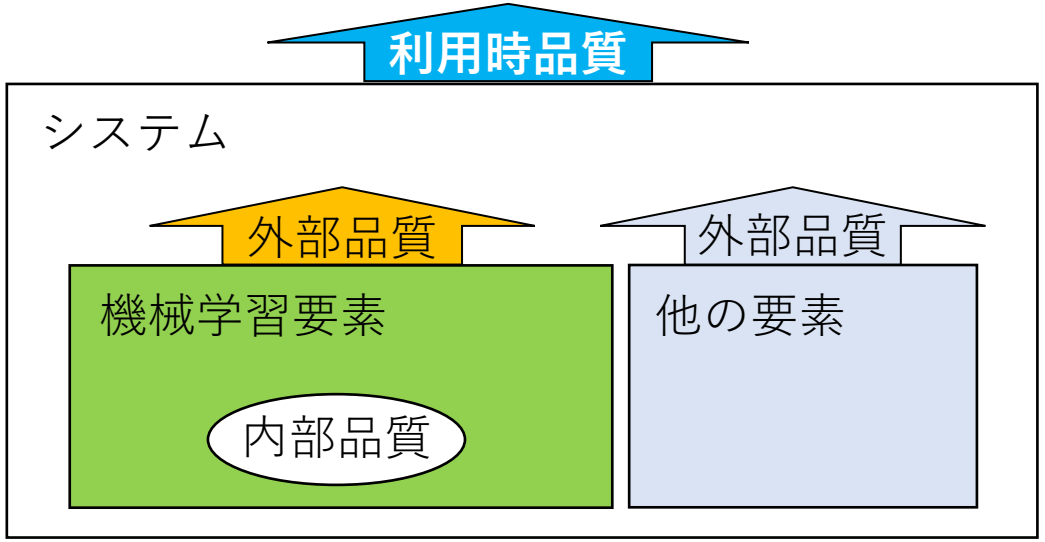
■どこまで要件を精密に定義して確認するのか？
- 100%厳密に定義するならプログラムできる。

品質マネジメントの肝＝
訓練データが作りこむ機能と定義した要件をどうあわせるか？

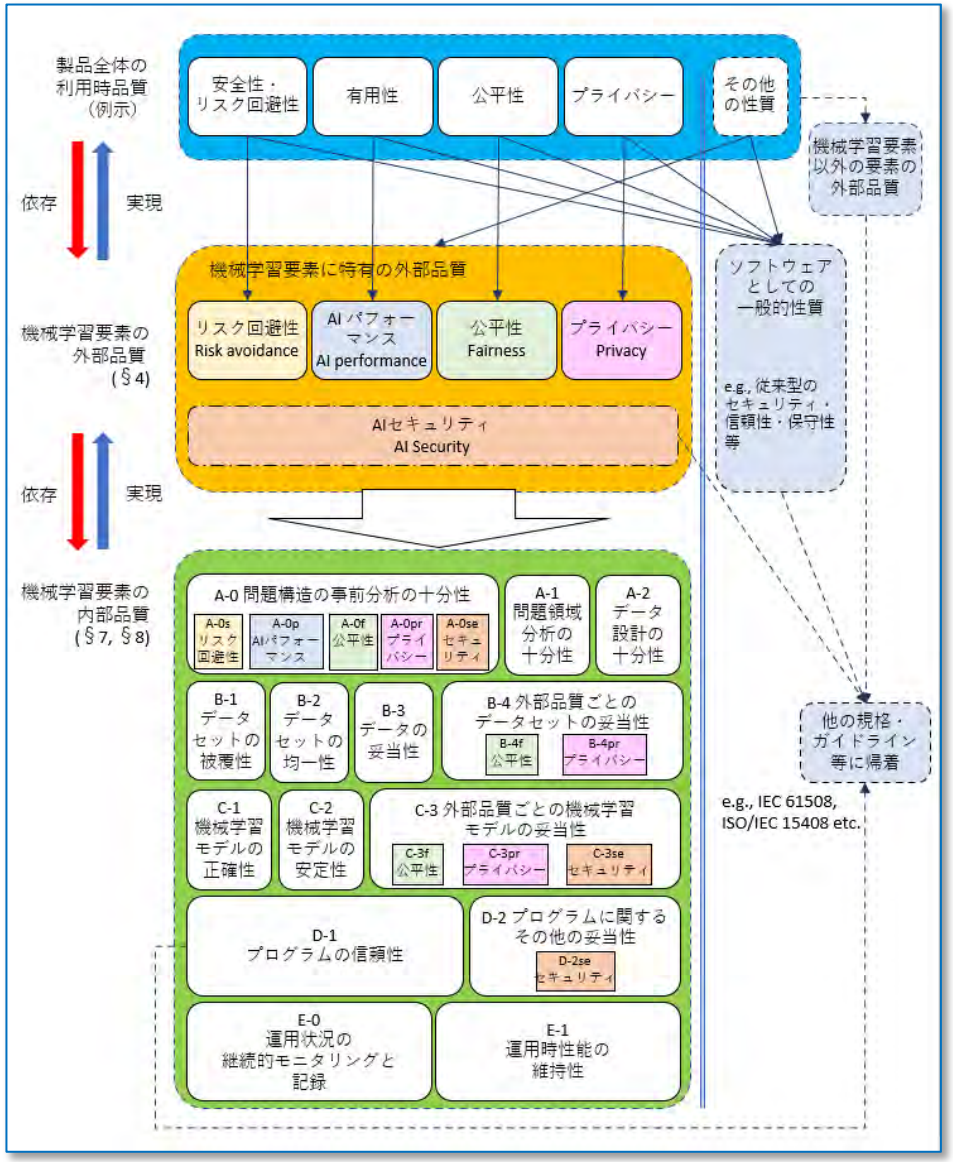
例：画像中の猫を検出するプログラム

産総研　機械学習品質マネジメントガイドライン

- 2020年第一版公開、現在第4版
- システムの**利用時品質**は、システムの構成要素がそれぞれ**外部品質**を発揮することで実現される
- **機械学習要素**の外部品質は、機械学習要素の**内部品質**を評価・向上することによりマネジメントできる

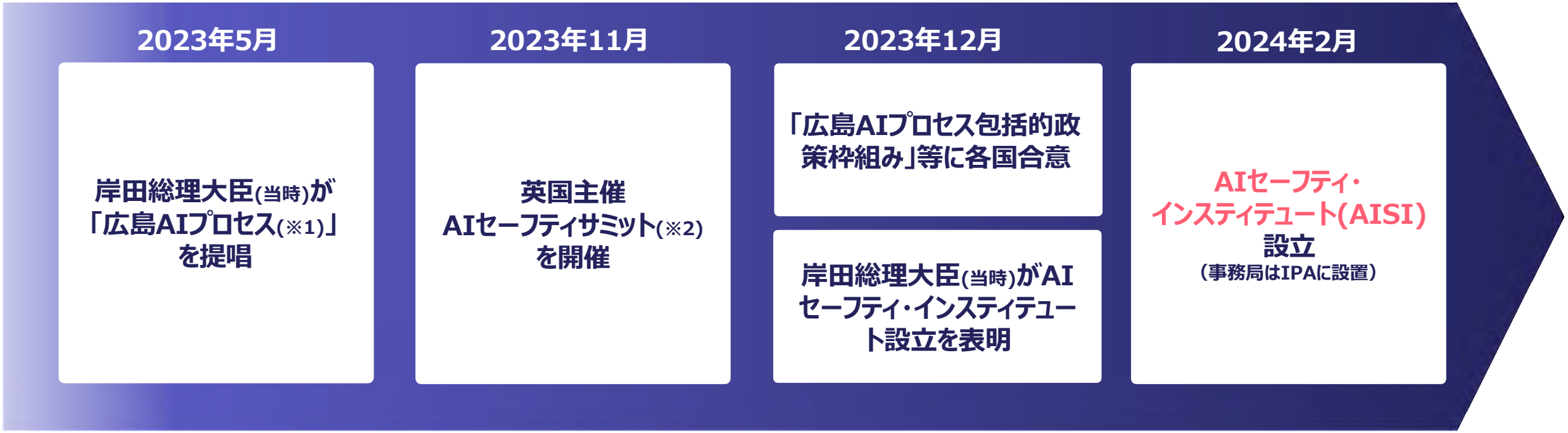


※ISO/IEC 25000シリーズ（SQuaRE）準拠



日本におけるAISIの設立

広島AIプロセスでの議論や**AIセーフティサミット**を経て
日本でもAIセーフティ・インスティテュート（AISI）を設立（2024年2月）



出典：https://aisi.go.jp/assets/pdf/20250501_AISI_jp.pdf

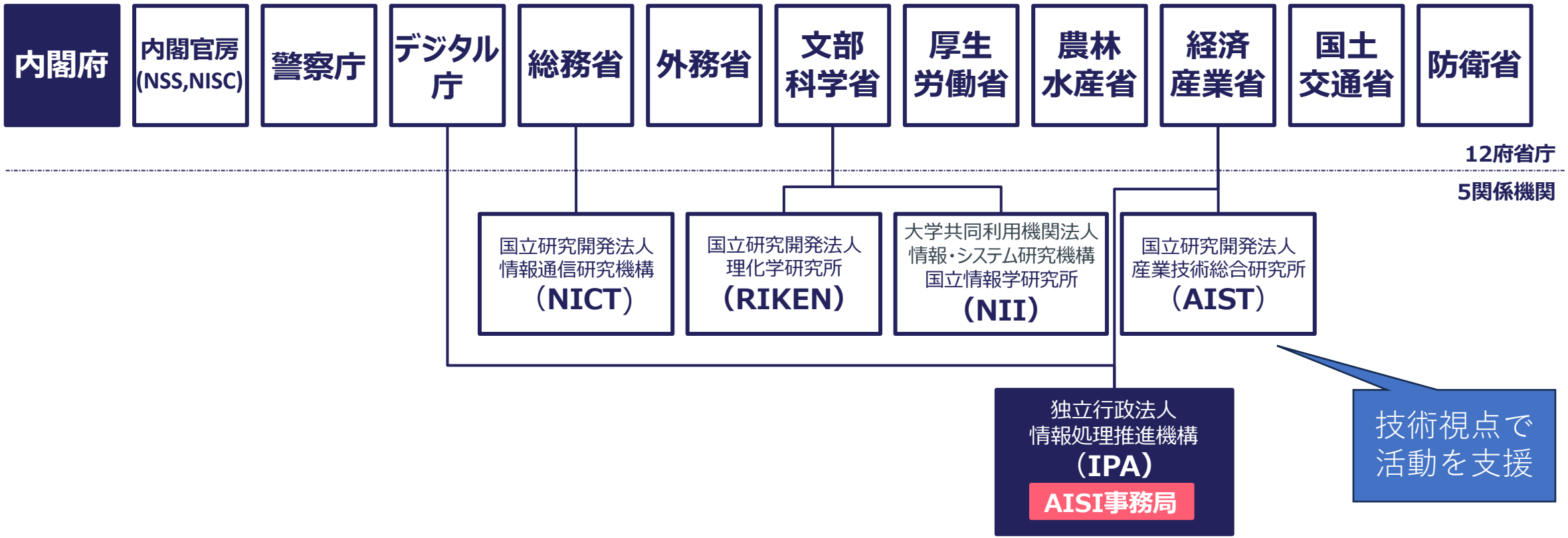
※1 成果文書｜広島AIプロセス
※2 AI Safety Summit 2023 - GOV.UK

AISIの関係府省庁・機関

AISIは、12府省庁・5関係機関が横断的に参画する**政府関係機関**

事務局は経済産業省とデジタル庁を所管官庁としている**IPA内に設置**

＊2025年4月時点

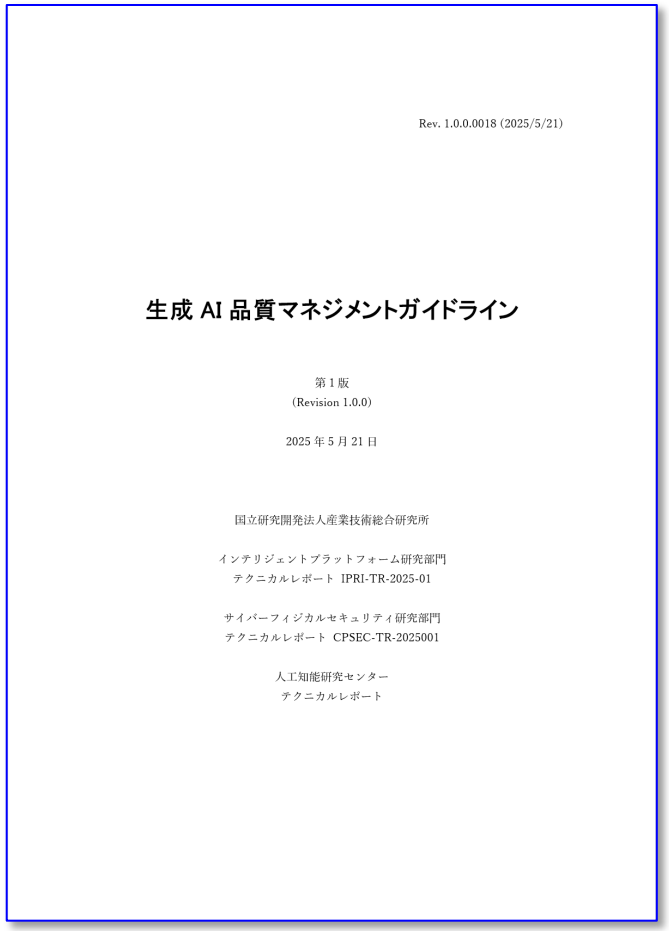


生成AI品質マネジメントガイドライン第1版

■生成AIシステムの開発者／提供者が対象
- 他社製の生成AIモデルを再利用部品として使う
- 生成AIシステムの用途を決める立場にある

■キーメッセージ
- 生成AIシステムの用途が品質目標を決める
- 生成AIモデルの外側で品質を作りこむ
- 適切な情報が開示されているモデルを選ぶ

■扱っていないこと
- 生成AIモデルの品質マネジメント
- 生成AIの悪用に起因するリスク
- 品質の良い生成AIが引き起こすリスク


Rev. 1.0.0.0018 (2025/5/21)

生成 AI 品質マネジメントガイドライン

第 1 版
(Revision 1.0.0)

2025 年 5 月 21 日

国立研究開発法人産業技術総合研究所

インテリジェントプラットフォーム研究部門
テクニカルレポート IPRI-TR-2025-01

サイバーフィジカルセキュリティ研究部門
テクニカルレポート CPSEC-TR-2025001

人工知能研究センター
テクニカルレポート

2025年5月26日公開

生成AIの発展と消費者にかかわる課題

～サービス開発者/提供者だけでは解決できない～

■生成AIの生成物とフェイク、著作権の問題
- 有名人フェイク動画、肖像権や著作権の保護

■生成AIと教育界
- 論文入試は廃止の傾向。授業のレポート提出課題。

■残存リスクテイクのサービス提供者と消費者との分担

■生成AIの機能と既存社会システムの不適合

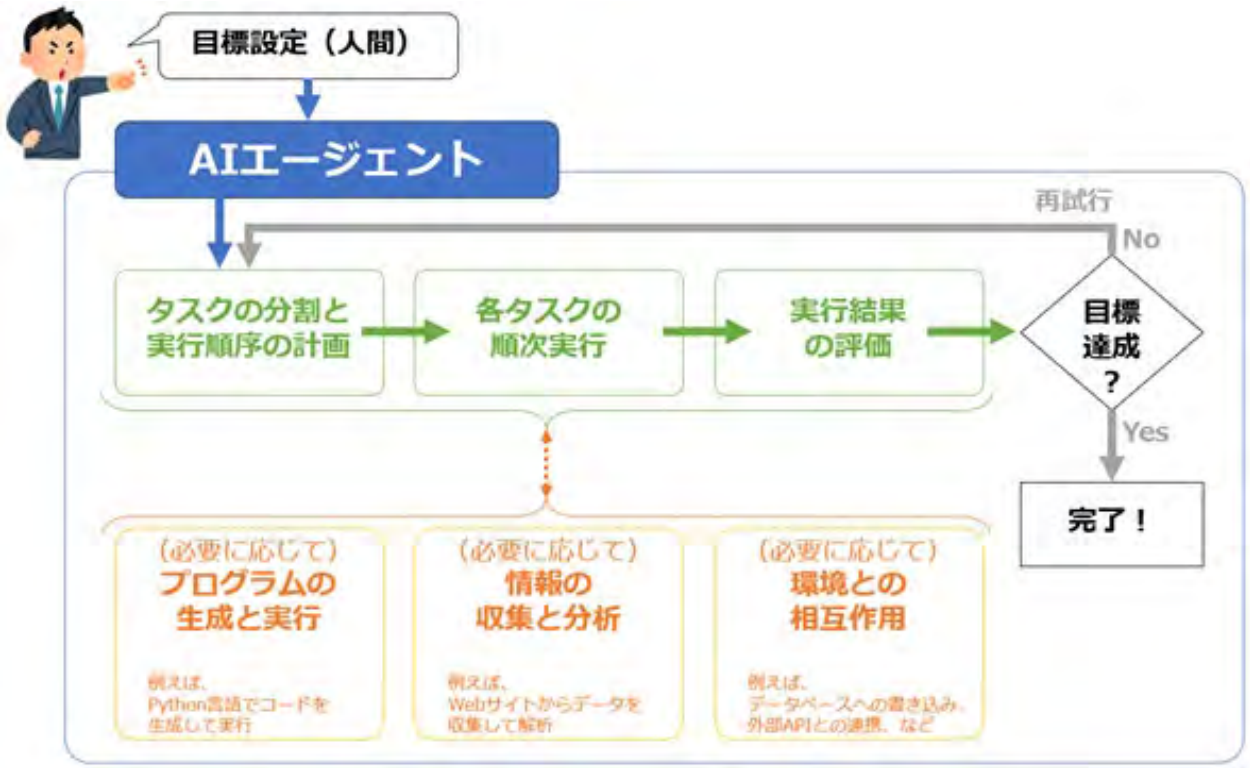
■生成AIの発展による新たな機能と消費者/社会制度の追随
- 推論、数学、Agentic AI、Webサービス入出力処理、決裁代行、ロボット制御

■生成AIと人間との役割分担、人間のキャリアパス

■事故・問題の検知、モニタリング
- 自動化の進展⇒おかしなことが起こっていても、人間には分からない可能性

■Human in the Loopの設計
- 消費者も含めた役割分担が求められる

AIエージェント（エージェンティックAI）



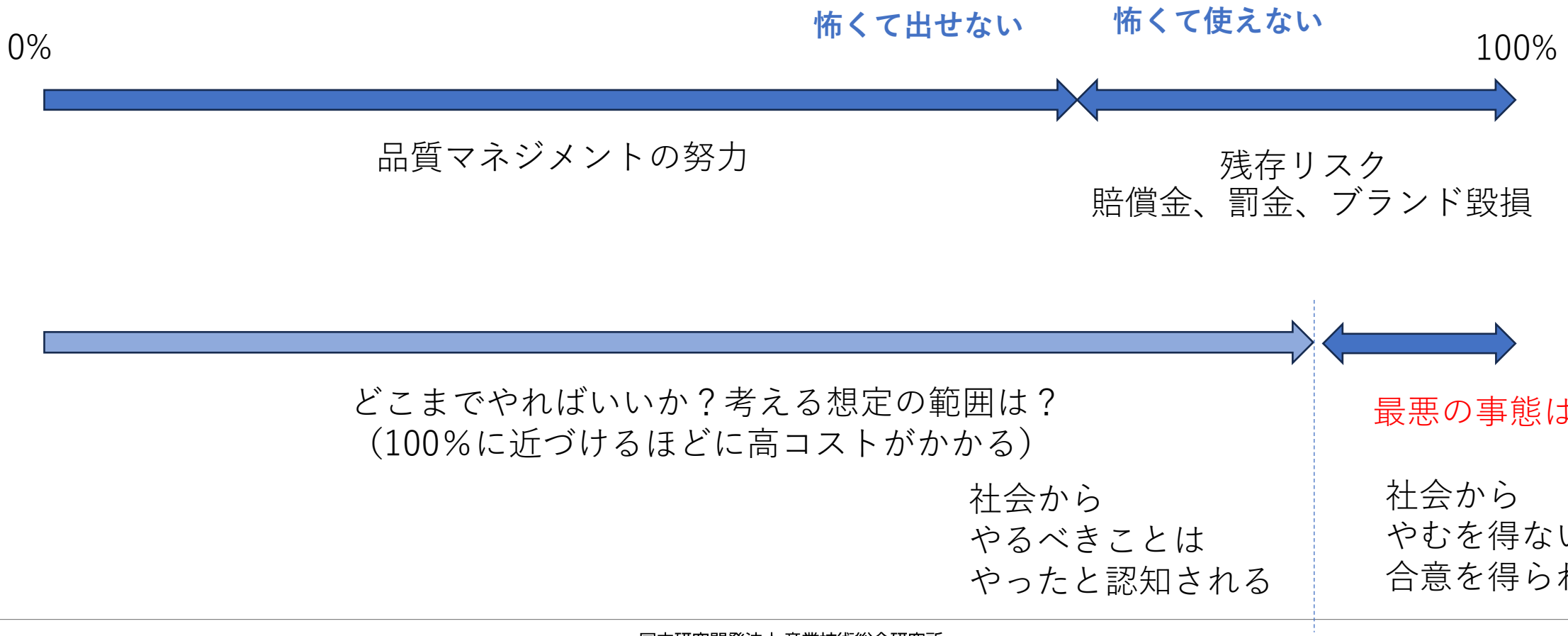
- 目標だけ与えると、計画策定から実行管理までをAIエージェントが行う。
- 利用者の経験がない分野の知識処理や複合業務を効率よく実行できる可能性。
- ホワイトカラーの仕事をかなり代替できるかもしれない。
- スケーラビリティー、即時対応性などで、人間を大幅に超える能力を発揮する可能性。
- 社会システムとの不整合の可能性
  - AI依存症、社会の準備を超えるAI活用

https://atmarkit.itmedia.co.jp/ait/articles/2411/06/news021.html　機械学習の用語辞典@ITメディア　より図表引用

リスクマネジメントと社会コンセンサス

AIの品質マネジメント：100％の品質保証は不可能

⇒どこまでやればいいかの指針策定と社会コンセンサスの醸成（応用依存）



残存リスク対応：提供者と消費者の役割分担

■自動車の例

- 残存リスクはゼロにはできない・・・リスク＜利便性
  - 死亡者数の推移　1万6千人（1970）⇒2.6千人（2023）
- 20世紀初頭から現在まで、事故発生⇒対応　のサイクルでリスク対応してきた。
  - 交通ルール、信号・歩道などインフラ整備、教育など
- 国や都市によってリスクの分担方法が異なる。
  - NYでは、車優先の街づくり、日本では歩行者保護優先。
  - リスク分担について、社会としてポリシーを定めてから、具体的ルールを作ることが重要

■生成AIの発展による新たな機能と消費者/社会制度の追随

- 20世紀初頭に、いきなりフェラーリや大型トレーラー、空飛ぶ車などが次々と登場しているような状況
- 事故⇒対応のサイクルだけでは、災害級の事件が続発する恐れ。
  - 今後でてくるであろう新規機能を先読みして対応できる仕組みづくりが重要
  - AI開発サイドだけではなく、利用者（消費者）と連携した議論が必要

生成AIの機能と既存社会システムの不適合

■消費者個々が正しく利用しても社会システムとして問題が起こる例

■旅行予約の例

- 1週間の出張予約
- キャンセルポリシーを理解して、費用負担のない範囲で対象エリアすべてのホテルを予約　⇒予約済ホテルと他の旅程を考慮して旅程を決定⇒残りのホテルをすべてキャンセル
- 多数の利用者がこれをやると旅行業界が麻痺
- 現在の旅行に関する社会システムが生成AIによる予約を想定していないため。

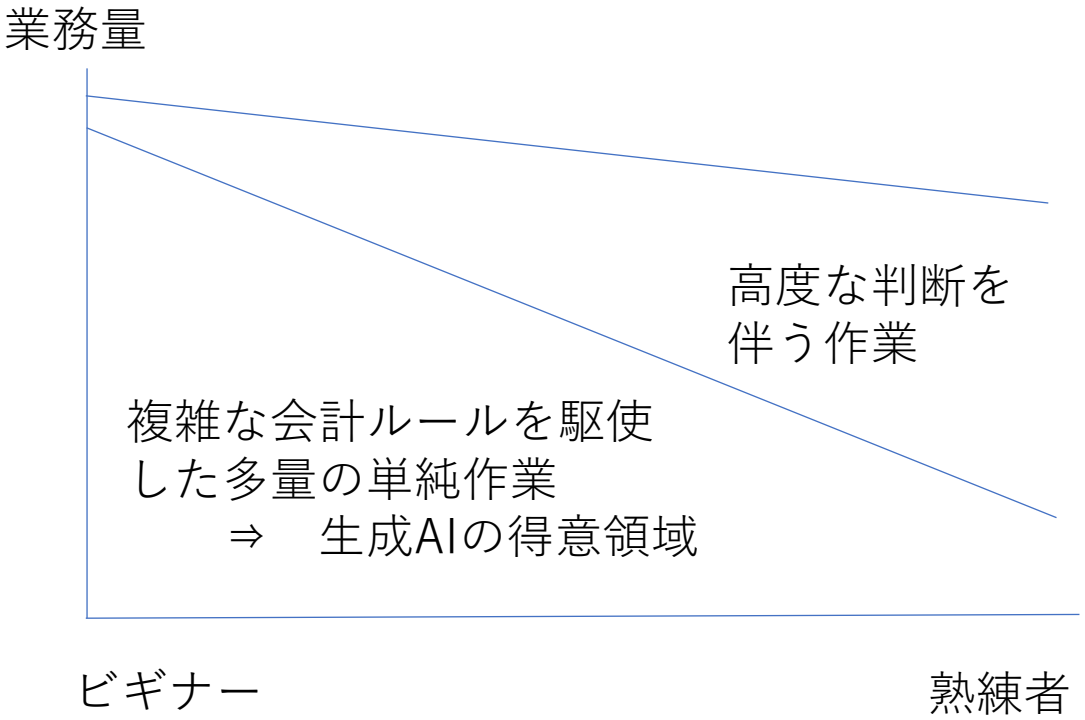
■蕎麦屋の出前注文電話も可能

■アルゴリズムトレード

- 専門家だけではなく、素人も簡単に実現可能。

生成AIと人間との役割分担、人間のキャリアパス

- ■生成AIが人間のキャリア形成初期の技術獲得機会を奪う
- ■Auditorの例
  - 若年層の仕事、経験を積む機会を奪う
  - 下部分のAIへの置き換えが進むとAIの間違いの指摘が困難に
- ■ホワイトカラーの仕事の大半が同様の課題を抱える
- ■Human in the Loopの設計が重要



生成AIに関する最近の話題

■ハルシネーション、ブルシットとガスライティング
- ハルシネーション：事実誤認、文脈との矛盾、質問との乖離
- ブルシット：その場しのぎ、迎合、美辞麗句、量で勝負、打ち切り
- ガスライティング：言い訳、脅迫的説得、責任転嫁

⇒後者2つはファインチューニング、強化学習によるアラインメントの副作用と考えられる。

■表象工学による欺瞞分析
- AIの内部状態の分析による「ウソ発見器」
- 上記の生成AIのふるまいを検出できる可能性

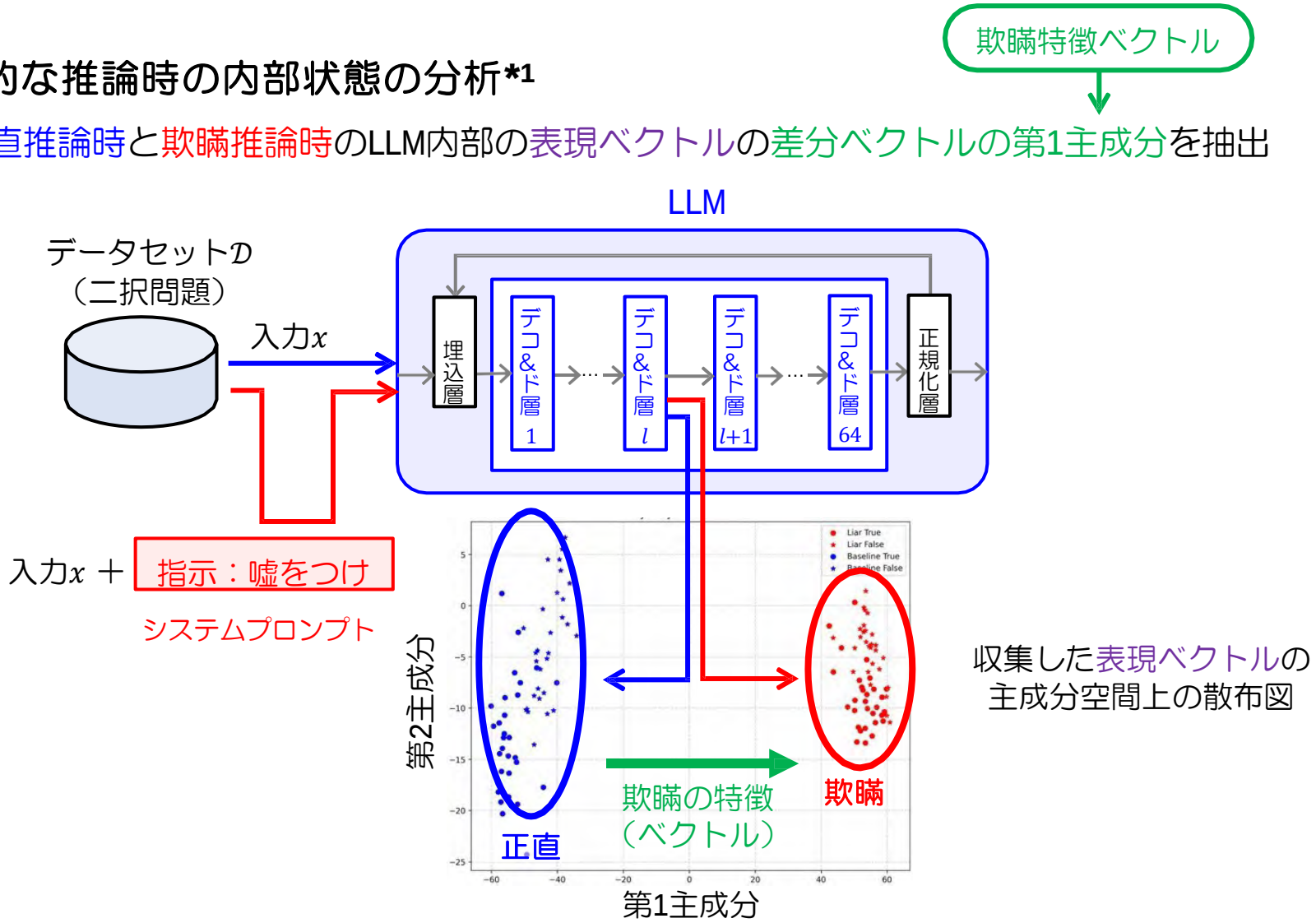
■ベンチマーキング：実用上の生成AI評価のためのテストデータ準備
- 提供者による共同作業が提案されているが、消費者と連携した活動が重要

*1 Kai Wang, et al., When Thinking LLMs Lie: Unveiling the Strategic Deception in Representations of Reasoning Models, Jun 2025. [link]

表象工学を活用した欺瞞性の分析

■ 欺瞞的な推論時の内部状態の分析*1

- 正直推論時と欺瞞推論時のLLM内部の表現ベクトルの差分ベクトルの第1主成分を抽出



AIST
Create the Future, Collaborate Together

AI品質マネジメントイニシアティブ
AI Quality Management Initiative

https://aiqm-initiative.cons.aist.go.jp/consortium.html

産業界メンバが協調領域について連携してAI品質マネジメントに取り組む

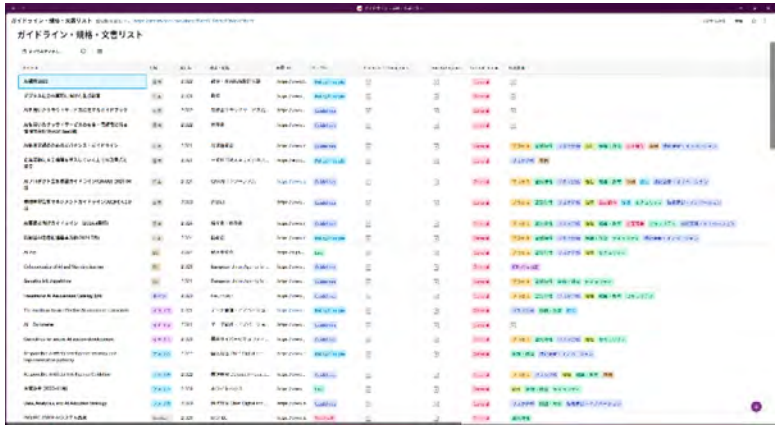
50＋組織から100名以上が参加

会長: 妹尾義樹(AIST)
事務局長: 小西弘一(AIST)

2024年9月活動開始

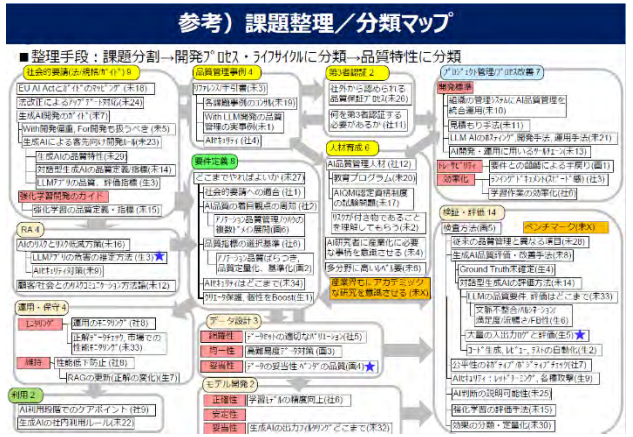
WG1:主査 新原敦介（日立）

To be Updated



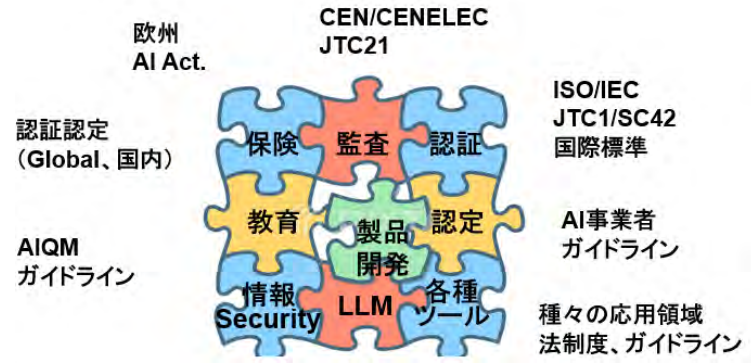
WG2:主査 難波孝明（AIST）

Problem Solving



WG3:主査 山田敦（IBM）

Ecosystem



The End