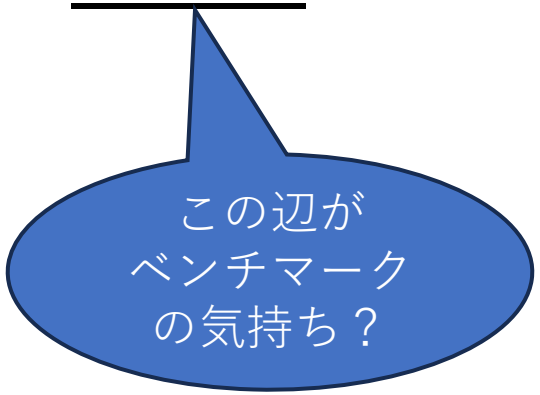
AIST
Create the Future, Collaborate Together

AISI Japan AI Safety Institute Partnership

機密性1情報　投影用

マルチモーダルAIのセーフティと共通基盤



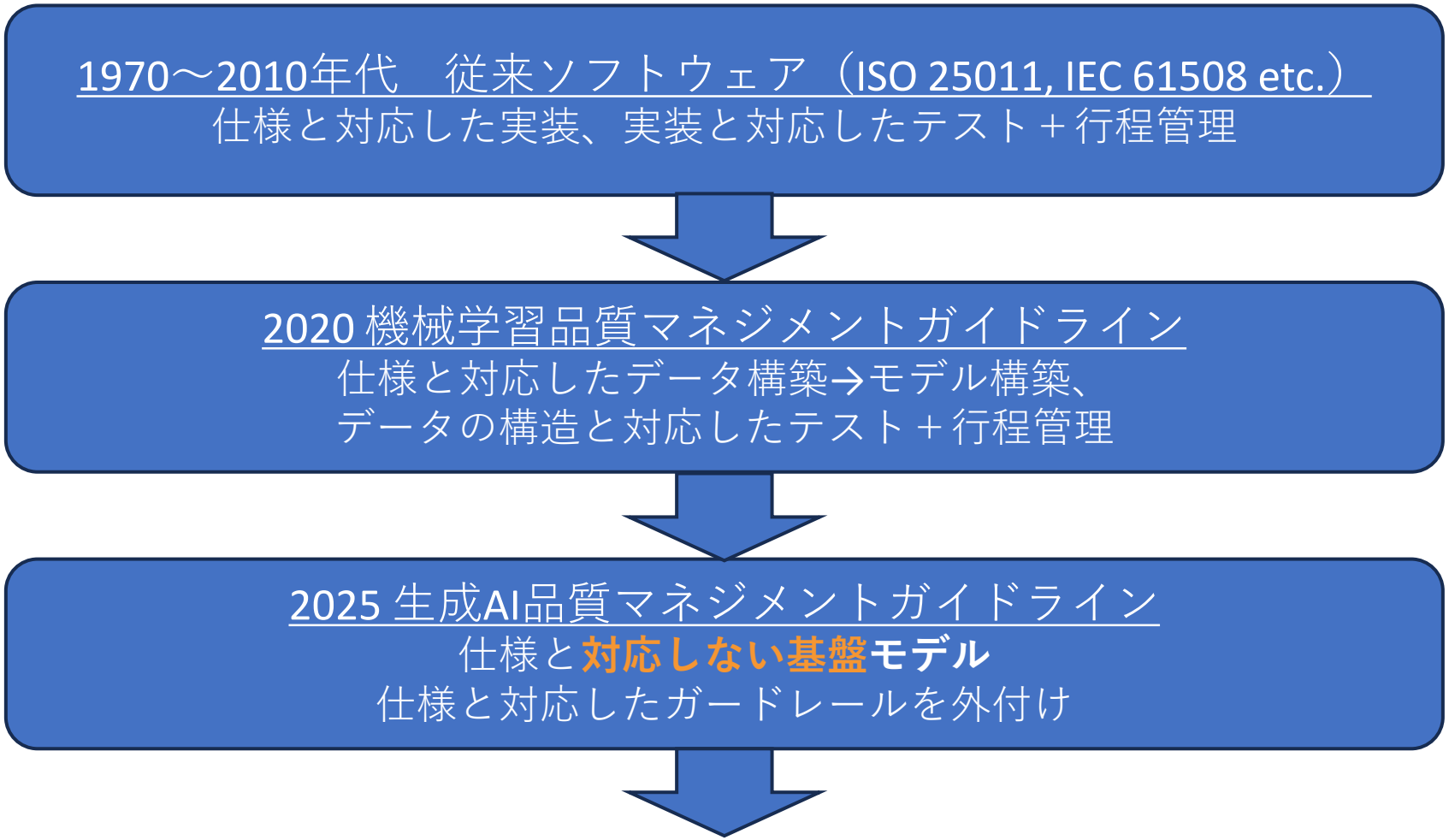
大岩　寛

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 副研究部門長
AI セーフティ・AISI パートナーシップ担当

2025 年 5 月 21日

AI の進化と品質マネジメント

- **AI の進化が著しい中で、システムづくりの構造も変わっている**



機械学習までの品質マネジメント

- **従来ソフトウェアの品質マネジメント（超ざっくり）**
  - **リスク分析 ⇒ 仕様と設計をきちんと固める（要求仕様・実装設計）**
  - **設計に沿ってシステムを作る ⇒ 設計や仕様に沿ってテストする**
    - 当然、テストそのものは一品ものになる（ファジングテストとかは除く）

- **機械学習の課題と品質マネジメント**
  - **実装と仕様の構造が全く一致しない**
  - **実装の代わりにデータを仕様から設計しよう**
  - **仕様に沿って「設計したデータで」テストする**
    - テストはやっぱり一品もの
    - ツールは共通化できた

機械学習
品質マネジメント
ガイドライン
（2020～2023）

生成AI・言語モデル（LLM）と品質管理・ベンチマーク

- LLM のシステム応用の典型的構成
  - LLM の性能と言語入出力能力に負うところが大きい
    ＝ システム内での LLM の存在感が大きい
  - 基本的に言語（プロンプト・応答・文書データ）を扱う
    ＝ 処理内容の「言語」の比重が応用対象のドメイン論理と比べても大きい
  - LLM の暴走を確実に止める手段は少ない ＆ LLM はほぼブラックボックス扱い
    ⇒ 外付けのガードレールでの制御が現実的

生成AI品質マネジメント
ガイドライン（2025）

- モデルが共通なら、一定のベーステストは共通にできないか？

⇒ 「ベンチマーク」の発想

LLM におけるベンチマークの意義

- **（おそらく）欧米の一部の視点**
  - **国の危機、あるいは人類存亡の危機を招く AI を検知したい**
    - AI の Capability の議論が結構熱い
    - 流通管理・国境管理の目的からベンチマークを国が持つ強い必要がある

- **開発者側の視点**
  - **共通のベーステストが欲しい**
  - **転移学習・ファインチューニングの前後での性能の劣化を検知しておきたい**
    - ベンチマークに含まれる、公平性など品質の提供そのものは**ベースモデル**の役割
    - でも、チューニング後にそれが保存されている保証はない

マルチモーダルモデルとベンチマーク

- 関根先生との最初の会話（注: 思い出しのでっち上げ）
  - 「マルチモーダルのベンチマークやってって言われたましたけど、
    やり方の計画とか決まってます？」
  - 「どうやればいいかまで含めて未知数、とりあえずよろしく」

- ということなので…
  - まずは問題整理ですね、と。

全体概要は
次の講演枠で

- 産総研側の AISI Partnership プロジェクト「AI セーフティ 1.0」
  - マルチモーダルAI 向けセーフティガイドラインの検討
  - 一旦引き取りました。

世にあるマルチモーダルモデルの作り方

- **典型的な作り方 3タイプ**
  - **Modular（モジュール型）**
    - 各モダリティを専用エンコーダで埋め込みに変換し、写像層を介して言語モデルに入力
    - 事前学習済みのエンコーダや言語モデルを凍結し、写像層のみを学習対象とする設計が基本
  - **Native（ネイティブ型）**
    - 複数モダリティを単一のトークン列として同一モデルで処理する統合構成
    - 深い相互作用を捉えやすいが、学習や運用に必要な資源が大きい
  - **Hybrid（ハイブリッド型）**
    - 構造はModularに近く、学習はNativeに寄せ、一体最適化の効果を取り込む

- **共通する特徴**
  - **言語モデルは既存のものがベース、別モダリティは応用適応が（現時点では）多い**
  - **表現空間の対応づけがシステム全体のふるまいに影響する**

世にあるマルチモーダルモデルの品質マネジメント

- **品質マネジメントの方針**
  - **For 現在のものづくりの方法**
    - LLM については、LLM 向けの既存ベンチマークなどを使う
    - 非言語モードについては、どちらかというと従来の機械学習に近い？
      - 特定のドメインに関するデータを流し込んでいる場合が多い
  - **まずはこちらに関する取り組み**

マルチモーダル AI 品質マネジメントガイドライン

- **2026年3月30日 第1版**
  - **主な対象**
    - 複数の異なるデータ形式を入力として受け取り、かつ出力として生成し、組み合わせて情報処理を行うAIシステム
    - 本書の主対象：テキストと静止画像を入力とし、テキストを出力する生成AI
  - モデル単体が複数モダリティを扱える場合に限らず、単一モダリティのモデルを複数組み合わせた構成も対象
  - **品質管理の対象 ＝ AIシステムの内部**
    - モデルの品質を管理 ＝ 機械学習に近い

- **第1章　はじめに**
  - 目的、対象範囲、想定読者、用語の定義
- **第2章　AI品質マネジメント**
  - 品質マネジメントの一般論の要点
- **第3章　マルチモーダルAI**
  - 前提とシステム構成、モデルアーキテクチャの類型
- **第4章　マルチモーダルAIの品質マネジメントの難しさ**
  - モダリティ間の照応、照応レベル、照応失敗の類型
- **第5章　マルチモーダルAIに特化した品質マネジメント手順**
  - 役割分担、データ設計、モデル品質、運用、評価、ガバナンス
- **付録1　適用事例による具体化**
  - EC商品説明生成、インフラ点検支援、SNSコンテンツモデレーション

品質の問題を分解・診断するための共通言語

産総研
ともに挑む。つぎを創る。

分析の枠組み：各モダリティの表現空間における、固有領域と共通領域を見分ける

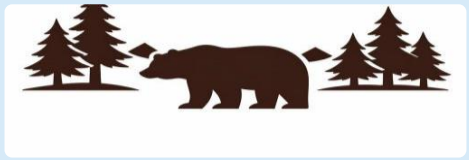
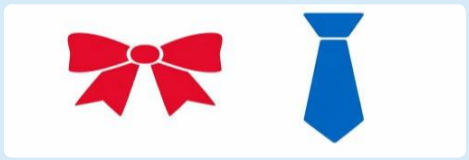
固有領域（モダリティ固有）
画像解像度・テキスト文法など各モダリティ固有の評価に委ねる

共通領域（照応）
モダリティ間の対応関係を照応レベル L1～L4 で段階的に分析

照応レベル L1～L4 — 対応の壊れ方を特定する座標系

本ガイドラインの核心

← 基本的な認識　　高度な認識 →

L1 対象の有無	L2 属性帰属	L3 事象間関係	L4 抽象的認識
			
「画像にクマが写っている」	「青いのはネクタイ」	「ネコがネズミを狙っている」	「お祭り」

注意：照応レベルの達成は品質の十分条件ではない
照応が正しくても、安全性の判断を誤る・公平性を欠く・不適切な断定をする問題はあり得る。照応レベルは「どの段階で壊れているか」を特定する分析軸

この座標系があれば、品質改善のために「何を直すべきか」を関係者が同じ言葉で議論できる

外部品質の達成をシステムとして担保する

産総研
ともに挑む。つぎを創る。

外部品質特性で品質目標を立てる

リスク回避性 | AIパフォーマンス | 公平性 | プライバシー | AIセキュリティ

品質の最終判定はこの5つの外部品質特性に対して行う

AIシステムの構成要素と品質達成の手段

実践の要

AIシステム全体

機械学習要素（MLモデル）
外部調達モデルを含む―内部への介入が困難な場合あり

周辺ソフトウェア ★品質達成の主要手段
プロンプト設計 ｜ RAG ｜ 出力フィルタ

モデルアーキテクチャごとに改善可能な範囲が異なる：

Modular型
各モダリティを独立MLで処理し連携 → 要素ごとに改善・評価可能

Native型
単一大規模モデルで全処理 → 内部介入困難、周辺SWが鍵

Hybrid型
両者の混合 → 改善可能範囲がアーキテクチャ依存

品質問題の切り分けと改善 ― 照応レベルを道具として使う

品質問題を発見 → 照応レベルで診断 → 原因の構成要素を特定 → 改善施策を策定

品質目標は従来の外部品質特性で立て、照応の分析は品質目標到達のための改善施策を考える道具として使う

共通基盤としての評価用道具立て

- 「開発者視点」の方の気持ちで

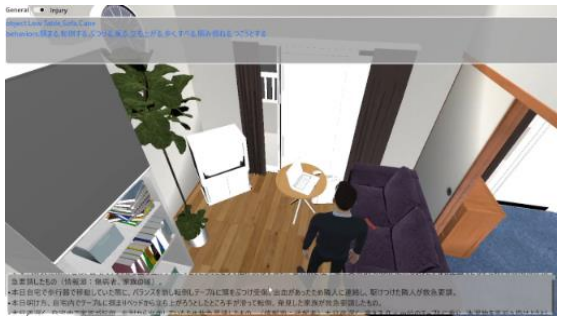
「なにかベンチマーク的に使えるものはつくれないか？」

⇒ ドメインごとの共通評価・開発基盤作成・公開

日常空間＋事故防止

- 生活事故状況DB
- 日常生活行動状況DB
- 仮想環境モデルデータ

⇒ **ありうる日常生活事故のシナリオを仮想環境で提示するシステム**

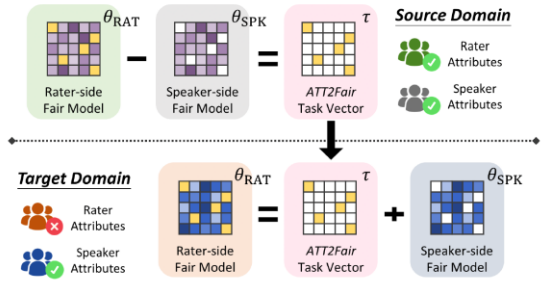


生活空間＋多様性

- ロボットシミュレーションフレームワーク
- 国内約50物件のフォトリアルな生活空間3Dデータセット
- 部屋の散らかり具合を調整できる物品配置

音声データ＋公平性

- 外国語訛り音声評価用データセットおよび音声AI学習プロセスの管理指針
- 音声モデルのバイアス除去手段



公共空間＋安全性

- お台場での屋内・屋外統合自律走行
- 運行管理SW ＋ 公共空間3Dデータ

マルチモーダル基盤モデルの品質マネジメント

- **マルチモーダル「基盤モデル」は来るか？**
  - 言語と比べても、モードの空間の「大きさ」が大きすぎる
    - 人間でも「物識り博士」はいても、全ての職業に万能な人は多分いない
- **基本的な空間識能力や安全性は、共通基盤化 できる可能性が高い**
  - 「安全を知っている基盤モデル」は強い
  - 産総研も開発を大いに狙っているところ
- **「基盤モデルを比較するベンチマーク」は必要になる**
- **応用寄りの性能に関しては、データでのベンチマーク共通化は難しいか？**
  - 「評価環境」など「**広義のベンチマーク**」は大いに期待できる
  - うまくいけば、応用ごとの評価環境はある程度共通化できる

- **そもそも汎用のマルチモーダルモデルは何を識っていないといけないのか？**

⇒ 後半？へつづく。

Create the Future, Collaborate Together