AIST Create the Future, Collaborate Together

AISI Japan AI Safety Institute Partnership

産総研のAIセーフティに対する取組み

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
AI セーフティ・AISI パートナーシップ担当

2026年10月16日 第2回AIセーフティシンポジウム

(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

科学哲学とAIと技術

- AI黎明期 ～ SF と恐怖の存在
  - HAL 9000 (?) / ターミネーター
    （サイボーグだけど「異質な存在」として）
- 2005年
  - 「シンギュラリティー」
- 2010年代後半
  - AI法規制・AI自主規制の議論
- 2020～2021年
  - AI を技術的に御す方向性
  - **哲学から工学技術へ**

社会からみたAIへの恐怖と要求: 前史

- **「人工知能」に対する感覚の相違**
  - **日本**
    - 多神教的・擬人化（人のようなもの）に比較的寛容
    - ドラえもん・鉄腕アトム など、敵にも味方にもなる存在
  - **欧米**
    - 感覚的には人と機械を峻別する傾向
    - HAL 9000 （2001年宇宙の旅）など、古くから恐怖の対象として描かれる
- **シンギュラリティー論（2005年）**
  - **AIの「自己進化」に対する恐怖感**

社会からみたAIへの恐怖と要求

- **2010年代後半**
  - **政治・統治のレベルに議論が（ようやく）到達**
    - 社会として認知して制御するレベルに至る
  - **要求を整理・明示して文書化する動き**
    - 社会原則・Principles レベル
    - 法律・社会的ガイドライン レベル
    - 技術ガイドライン・技術標準レベル

- **2019～2021年～……**
  - **技術としてAIを具体的に制御する流れ**
    - やっと、「得体の知れたもの」として扱うように

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

EU AI act
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

産総研: 機械学習品質マネジメントガイドライン （2020）

QA4AI コンソーシアム（2020）、Franhofer のガイドラインなど実装のガイドラインが相次いで発表される

→ **ISO TR 5469: Functional Safety and Artificial Intelligence (2024)**

2022年頃までの流れ

- AI は得体の知れないもの、という時代
- 何とか、技術として工学で御せるようにしてきた
- 結果として、産業標準化の対象として扱うようになった

……はずだった。

科学哲学とAIと技術

- AI黎明期 ～ SF と恐怖の存在
  - HAL 9000 (?) / ターミネーター
    （サイボーグだけど「異質な存在」として）
- 2005年
  - 「シンギュラリティー」
- 2010年代後半
  - AI法規制・AI自主規制の議論
- 2020～2021年
  - AI を技術的に御す方向性
  - **哲学から工学技術へ**
- **2022年**
  - **ChatGPT ショック**
  - **科学哲学に一気に逆戻り**

ChatGPT ショック

- **2022年11月のChatGPT公開**

- **AI が「得体の知れないもの」に戻った**
  - **世界的なルール形成の動きの加速**
    - 2023年5月 G7 広島 AI process
  - **各地域での AI Security Institute の設立と横連携**
    - US AI Safety Institute ▶ **US Center for AI Standards and Innovation**
    - AI セーフティ・インスティテュート
    - UK AI Safety Institute ▶ **UK AI Security Institute**

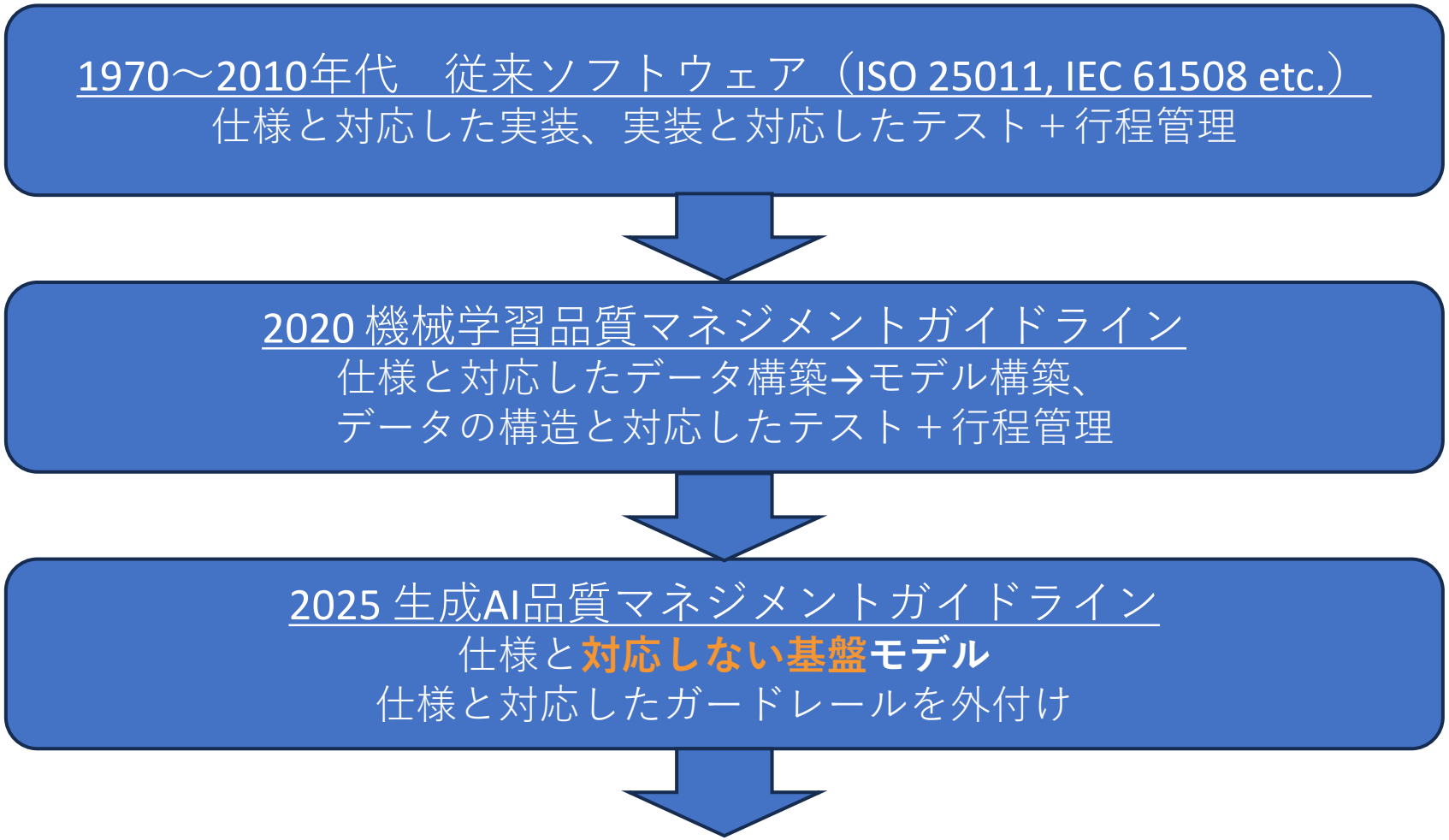
AIセーフティ強化に関する研究開発プロジェクト

- **2025年度より新たな研究体制を構築**
  - **セーフティの専門家とAI専門家の更なる協働体制の立ち上げ**
  - **研究領域の拡大**
    - 従来からのリスク管理・マネジメントの技術やガイドライン
    - **基盤AIなどのAIシステムそのものの開発も「セーフティ」の取り組みの一部に統合**
    - **応用領域の具体的なリスク分析やデータ整備**
  - **研究対象の拡大**
    - 基盤AI・言語モデル
    - 画像・音声などのマルチモーダルAI
    - フィジカル領域（ロボット・生活空間・道路・公共空間）のAI
    - いわゆる「エージェンティックAI」

(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

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**



(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

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）

(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

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）

(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

マルチモーダルAI と AIセーフティ

- マルチモーダルAI: 複数のモード (mode, modality) にまたがる処理・判断を行う人工知能
  - 典型的には、「言語による指示 + 画像・音声・動画」
    または「画像・音声・動画 + 言語の空間での意味理解」

- ロボットの世界と人工「知能」の融合 = 賢いロボット
  - ロボットの眼・耳としての機械学習AI （+ロジックベースの判断）から、
    ロボットの脳としてのマルチモーダルAI へ

(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

安全性担保の実装「粒度」の問題

- **機械／ロボット／whatever の安全をどう作り込むか？**
  - **人間でも2つのロジックが共存**

    ① **特定の作業における安全な作業手順の訓練**: 車の運転免許とか、工作機械の制御

    ② 人間の基礎としての共通**安全判断・生存本能による危険回避**

- **安全ロジックの作り込み**
  - **従来ソフトや言語（LLM）と比べても問題が大きすぎる**
    - 究極的には、「この生活空間の安全」「この世の安全」を判断させないといけない
  - **この世の全ての危険な状況の判断を教育させるのも茨の道**
  - **かといって、ワーク毎に出会う危険な状況を洗い出して個別に教えるのも茨の道**

(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

基盤モデル・共通モデルは有り得るか？

- 個別応用ごとのモデルは、安全論証はやりやすい
- 一方で、開発にもなかなか限界がある
  - リスク分析の深掘りの限界
  - （データを集めるコスト × 開発回数）の膨大さ

- 何か共通化が出来ないか？
  - LLM は（言語以外のマルチモーダル応用でも）有り難い。
  - これのフィジカル版、は有り得るか？

(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

基盤モデル・共通モデルは有り得るか？

- デメリット ＝ 想定シナリオの爆発
  - 「実社会のありとあらゆる安全判断」なんてものを分析しきれる、というのは絵空事
  - 同じ状況でも違う判断が正解、というのは応用によっては有り得る

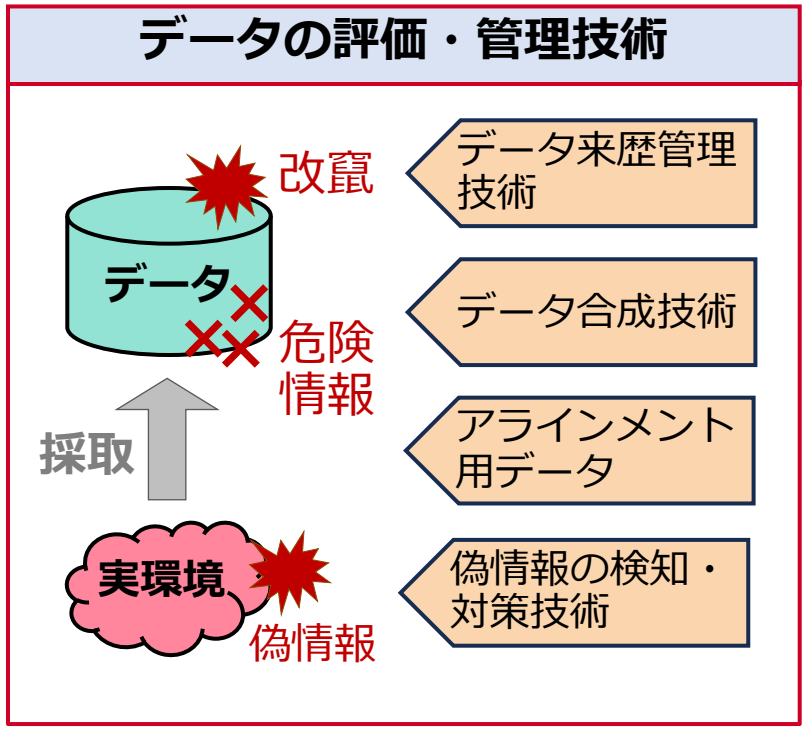
- メリット ＝ 開発の練度 と 規模
  - 1回だけの開発でよければ、最高の人材と大規模な投資を投入できる
  - 結果として、分割した個々の開発より安全になることは十分有り得る

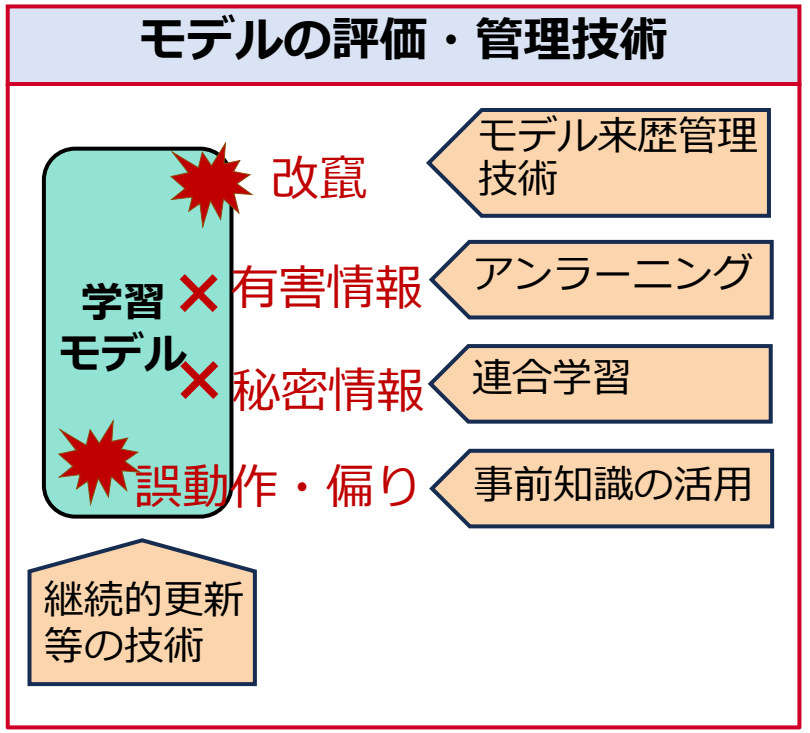
  - ソフトウェア工学において「再利用性」は重要な概念
  - まさにLLM が成功例、二匹目を狙いたいところ

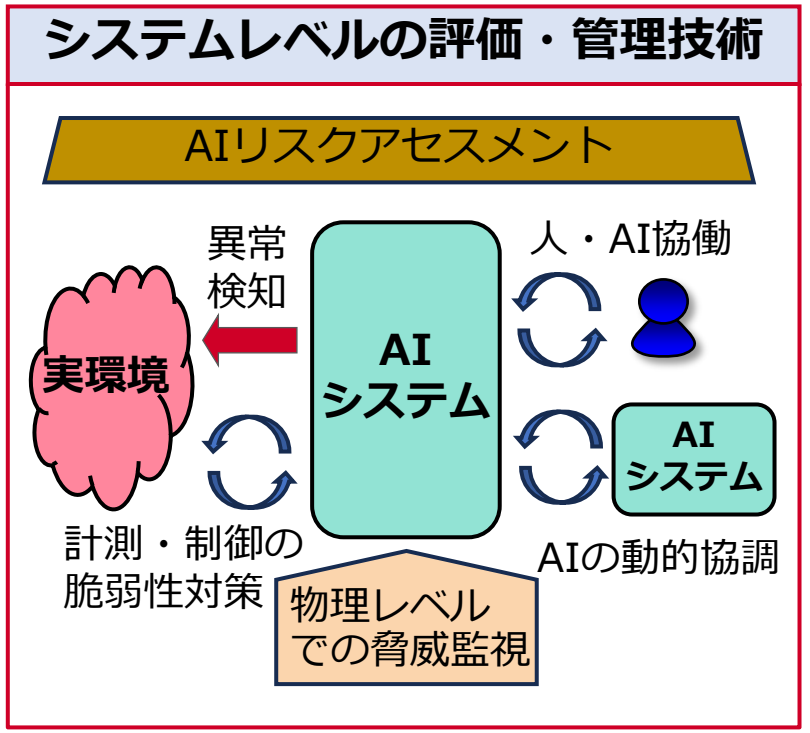
(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

• AIセーフティー基準の基盤となる評価・実装技術の確立

- ①AIセーフティー基準を技術的に詳細化し、②応用領域別の評価・実装技術に用いる要素技術を開発
- データ、モデル、システム等のAI資産ごとに、AIセーフティーの評価・実装技術を開発
- AIシステムライフサイクル全体でのリスクアセスメント技術を開発







(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

② 応用領域別AI セーフティ評価・実装技術

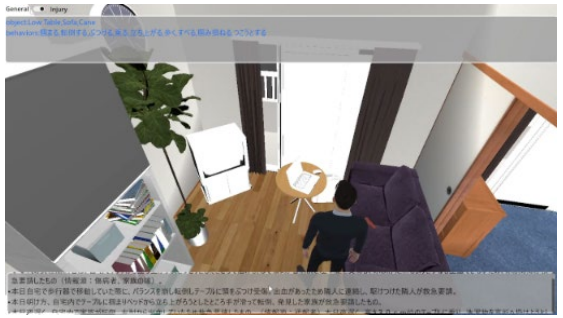
- 「なにか共通に使える、今作れるものはないか？」

⇒ 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**



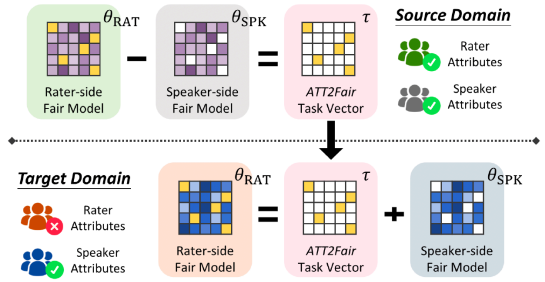
生活空間＋多様性

- 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 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

(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

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？**

(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

14:25-14:45 (20分)	**AIと意見が割れた。そのとき人は正しく意思決定できるのか？ ～知覚精度・Trust Calibration・Human-AI Teamingから考えるAIセーフティ～** 野里 博和（国立研究開発法人 産業技術総合研究所 人工知能研究センター 人間AI協働機構研究チーム 研究チーム長）
14:45-15:05 (20分)	**生活空間における見守りAIの普及に向けて －リビングラボを活用した安全性評価基盤の構築と社会実装へ取り組み－** 宮田 なつき（国立研究開発法人 産業技術総合研究所 人工知能研究センター 生活行動モデリング研究チーム 研究チーム長）
15:05-15:20 (15分)	**Coffee break**
15:20-15:35 (15分)	**人に接触し、身体を支え・動かす ロボットの安全** 藤原 清司（国立研究開発法人 産業技術総合研究所 情報・人間工学領域 研究チーム長）
15:35-15:50 (15分)	**VLAへのセンサ攻撃を対象としたセーフティコストによる定量評価** 飯島 涼（国立研究開発法人 産業技術総合研究所 サイバーフィジカルセキュリティ研究部門 研究員） 松本 勉（国立研究開発法人 産業技術総合研究所 サイバーフィジカルセキュリティ研究部門 フェロー）
15:50-16:05 (15分)	**データ拡張によるクリーンな少量データからの高信頼モデルの構築と評価** 山崎 啓介（国立研究開発法人 産業技術総合研究所 人工知能研究センター 機械学習研究チーム 研究チーム長）
16:05-16:20 (15分)	**科学のための基盤モデル・言語モデル・自律実験とそれらの連携** 椿 真史（国立研究開発法人 産業技術総合研究所 人工知能研究センター 主任研究員）
16:35-16:50 (15分)	**事例で見るAIセーフティと活用の両立** 杉山 阿聖 氏（株式会社Citadel AI リサーチリード）
16:50-17:05 (15分)	**Auditability of AI pipelines** 藤井 光 氏（株式会社コーピー CEOオフィス プログラムマネジャー）
17:05-17:20 (15分)	**行動するAIの品質マネジメントに向けて** 小西 弘一（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 総括研究主幹）

(c) National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.



Create the Future, Collaborate Together