行動するAIの品質マネジメントに向けて

前提の変遷と今後の展望

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

2026年10月16日

要旨

AIの品質マネジメントには、「**AIに何をさせたいのか**」をまず明らかにする必要がある

認識・予測AIでは比較的容易だった。

生成AIになって難しくなった。

「行動するAI」（AIエージェントやフィジカルAI）では更に難しくなる恐れがある。

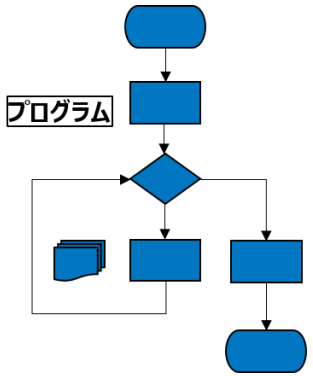
本講演ではその難しさを例示し、その対策に向けた”見立て”をお話する。

本講演のスコープ

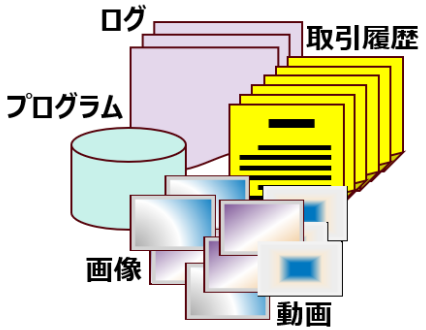
産総研
ともに挑む。つぎを創る。

従来ソフトウェアから始め、AIエージェント、フィジカルAIまでを展望する

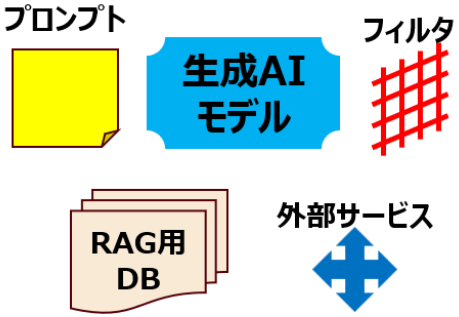
従来ソフトウェア



従来AI



生成AI
（含マルチモーダル）



AIエージェント



フィジカルAI



機械学習AIの品質の必要性

機械学習AIの用途例

自動操縦
- 自動運転
- 産業用ロボット
- ドローン

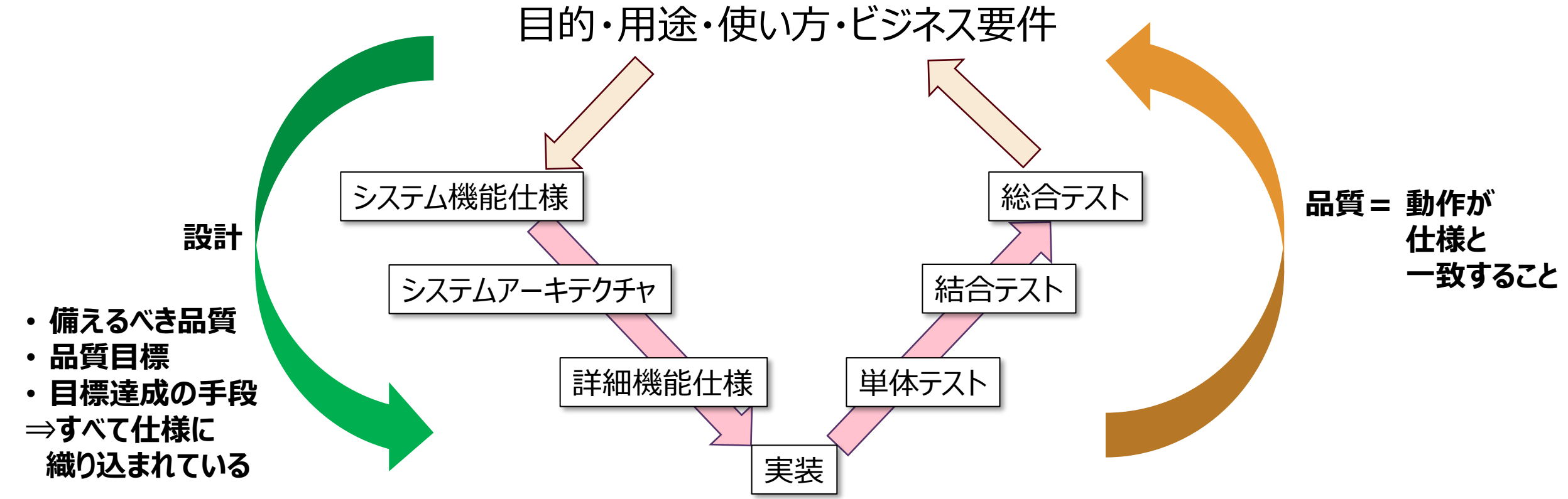
人に関する判断の支援
- 与信審査
- 採用判定
- 再犯リスク予測

機械学習AIに伴うリスク



従来ソフトウェアの品質マネジメント

作ったものをテストして動作が仕様と一致することを確認する
- 前提：設計が導出する仕様は正しい（正しくなるまで修正できる）

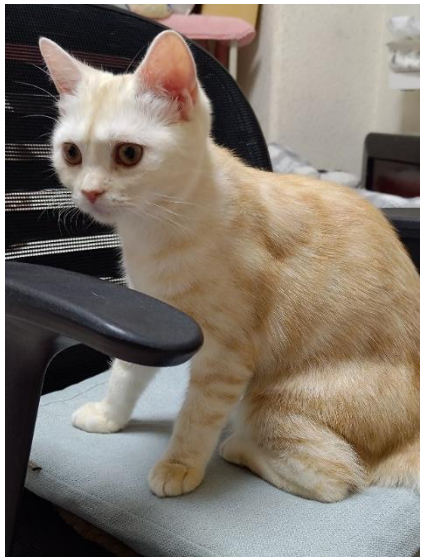


- **備えるべき品質**
- **品質目標**
- **目標達成の手段**

⇒すべて仕様に織り込まれている

機械学習AIの品質マネジメントの課題

- 機械学習で作るプログラムには仕様がない
  - 仕様が書けない“猫問題“に対処するためにこそ機械学習を使う



⇒動作と仕様を比べるテストの代わりに
何を測定し評価すればよいか？

⇒何かを測定し評価したとして、
それがどうなっていれば高品質といえるのか



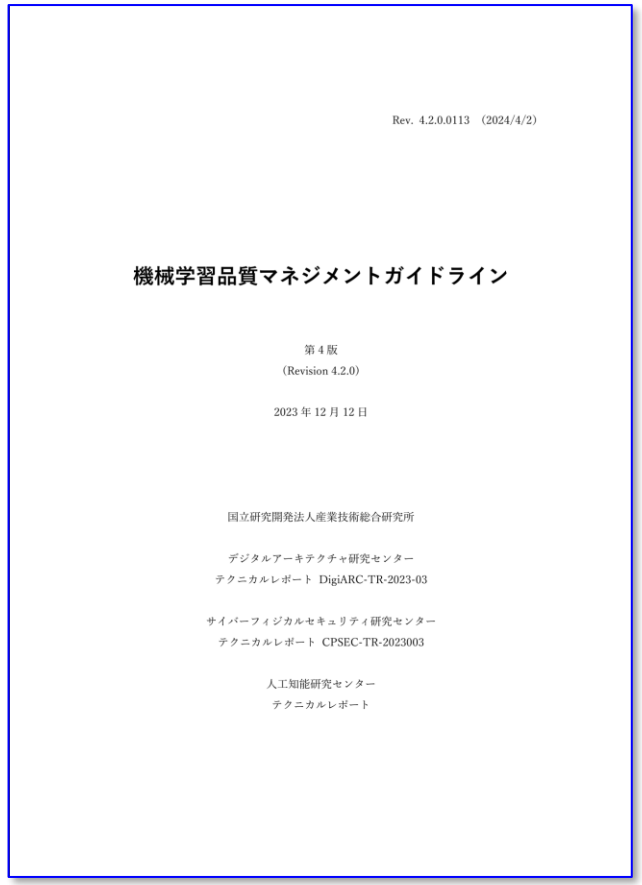
機械学習品質マネジメントガイドライン

産総研
ともに挑む。つぎを創る。

従来AI向け品質マネジメント手法

機械学習AIの品質を「作り込み」「確認し」「説明する」ためのガイドライン

- 主な想定読者
  - 機械学習を利用して作られる**製品やサービスの提供者**
  - 実際に製品・サービスをソフトウェアとして実装する**システム開発者**
- 2次的な想定読者
  - 製品やサービスの**利用者**：製品やサービスを選択する基準として
  - 第三者**評価機関**：品質評価・認証の基準として


Rev. 4.2.0.0113 (2024/4/2)

機械学習品質マネジメントガイドライン

第４版
(Revision 4.2.0)

2023 年 12 月 12 日

国立研究開発法人産業技術総合研究所

デジタルアーキテクチャ研究センター
テクニカルレポート DigiARC-TR-2023-03

サイバーフィジカルセキュリティ研究センター
テクニカルレポート CPSEC-TR-2023003

人工知能研究センター
テクニカルレポート

2023年12月12日第4版公開

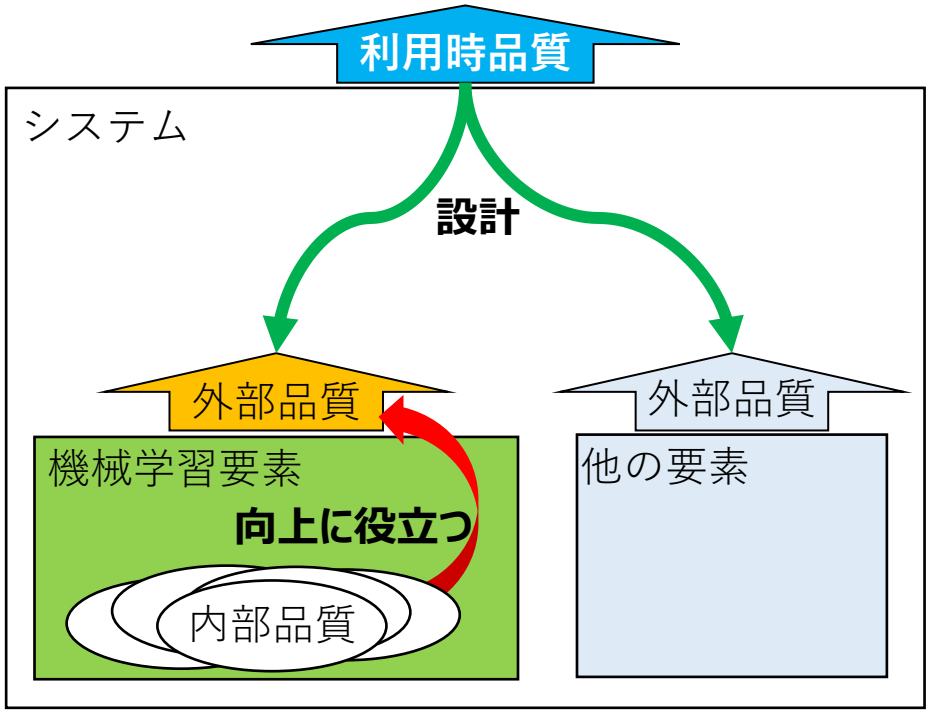
https://qai.pj.aist.go.jp/guidelines/machine-learning/index.html

MLQMガイドラインが提案する手法

従来AI向け品質マネジメント手法

- AIシステムをどう使って何をさせたいかを明示せよ
- AIシステムが備えるべき品質を定めよ

- AIモデルが備えるべき品質を設計によって導け
- AIモデルの開発に使うデータを設計せよ
- 訓練したモデルを評価して、必要な品質を満たせ
- 運用開始後もモニタリングして、問題が出たら対処せよ



システム品質をマネジメント対象にする

- 生成AIモデルには決まった用途がないので品質目標を決められない

なので

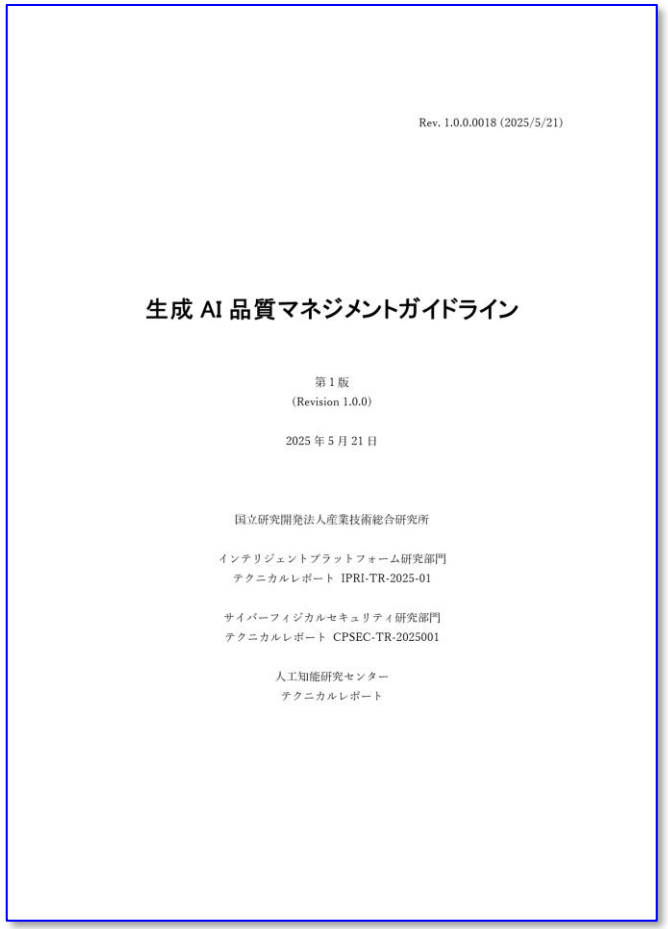
- モデルを部品とするシステムの用途に基づいて品質目標を定める
  - 幼い子供が使うシステムでは不快・危険な出力は絶対に阻止するべき
  - パーティの余興に使うのなら不正確な情報もきわどいジョークも問題にならない
  - 警察が捜査の参考にするなら武器製造法など危険な情報がむしろ必要かもしれない

生成AI品質マネジメントガイドライン第1版

産総研
ともに挑む。つぎを創る。

生成AI向け品質マネジメント手法

- 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

https://qai.pj.aist.go.jp/guidelines/generative-ai/index.html

2025年5月26日公開

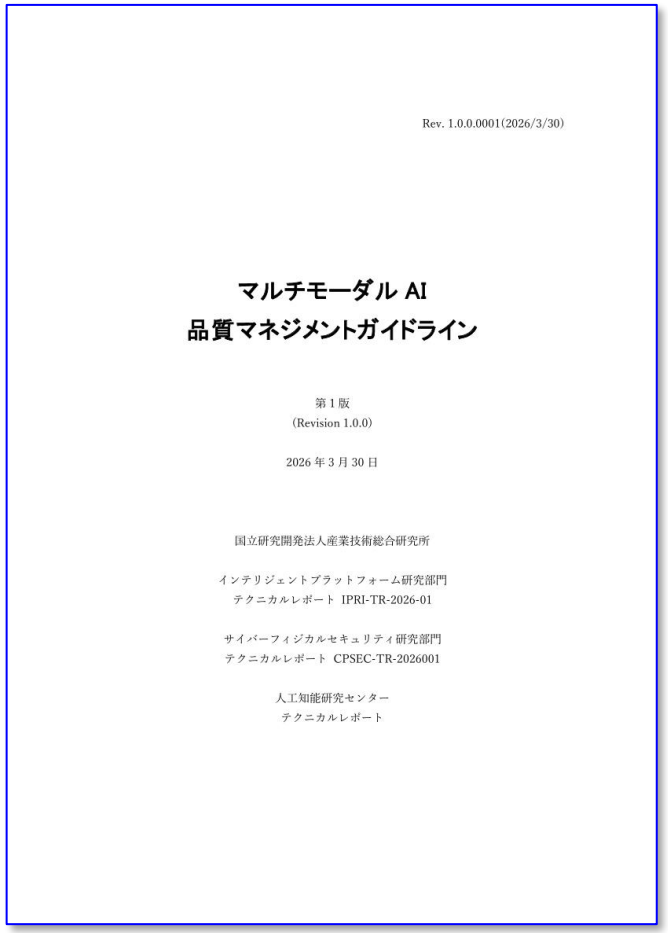
マルチモーダルAI品質マネジメントガイドライン第1版

産総研
ともに挑む。つぎを創る。

マルチモーダルAI向け品質マネジメント手法

- マルチモーダルAIシステムの開発者／提供者が対象
  - 生成AIシステムの一種としてのマルチモーダルAIシステムを取り上げている
  - この版では、テキストと画像を入力としてテキストを出力するシステムを想定

- キーメッセージ
  - **クロスモーダル照応能力**を考慮する必要がある
  - システムに求められる要件そのものではない
  - システムの利用時品質を実現するための前提条件


Rev. 1.0.0.0001(2026/3/30)

マルチモーダル AI
品質マネジメントガイドライン

第 1 版
(Revision 1.0.0)

2026 年 3 月 30 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-2026-01

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-2026001

人工知能研究センター
テクニカルレポート

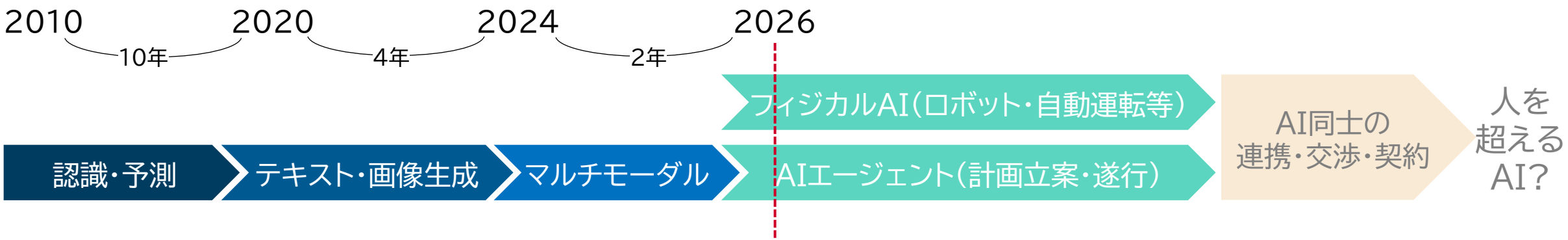
https://qai.pj.aist.go.jp/guidelines/multimodal-ai/index.html

2026年3月30日公開

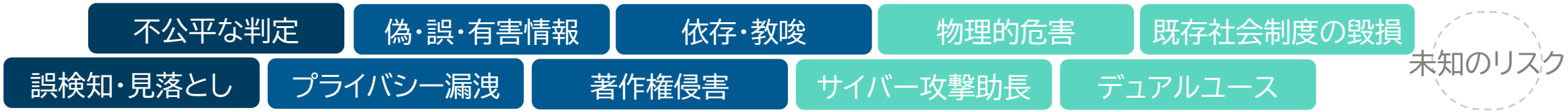
今後の展望

急速に発展するAI技術が新たなリスクを引き起こしている

AI技術の急速な進展



新たなリスクの顕在化



⇒技術が実現してからAIセーフティ規準を作るのでは間に合わない

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

Medium-Term AI Risk Foresight（MTARF: 中期的AIリスク展望）

産総研
ともに挑む。つぎを創る。

今後の展望

- 5-7年先のAIリスクを予想し、それに応じた研究課題を提示
- AIリスク具現化フレームワーク
  - AIリスクの源泉から危害までの連鎖を類型化して示したもの
- 中期的研究課題
  - AIの能力の拡大、地域・文化の違い、対策手段の広がり、前提の崩壊などを踏まえた課題15項目を提示



Medium-Term AI Risk Foresight

March 31, 2026
(Revision 1.0.0)

Committee for Machine Learning Quality Management

Artificial Intelligence Research Center
Technical Report
Cyber-Physical Security Research Institute
Technical Report CPSEC-TR-2026002
Intelligent Platform Research Institute
Technical Report IPRI-TR-2026-02

National Institute of Advanced Industrial Science and Technology (AIST)

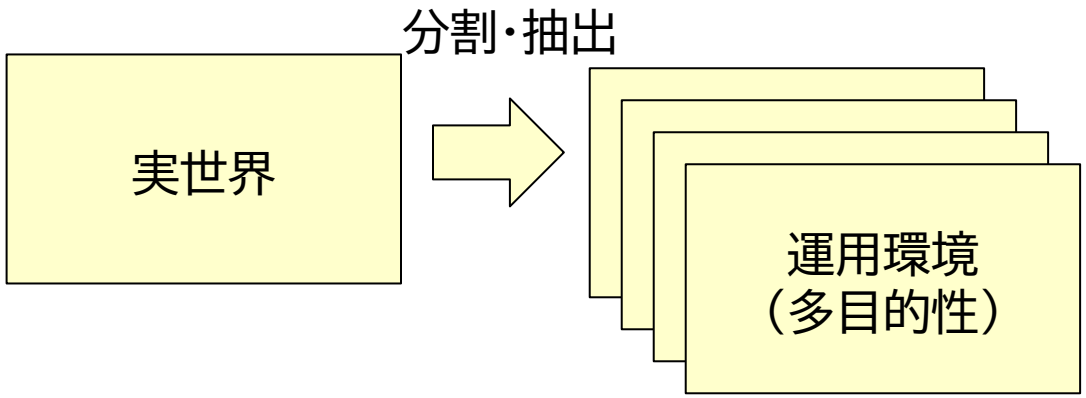
https://qai.pj.aist.go.jp/aisafety1/results/2025/3-1-1/pdf/Medium-Term_AI_Risk_Foresight-rev1.0.0.pdf

2026年3月30日公開

AIの行動範囲の拡大に伴う課題1（未来のAIロボットを例に）

今後の展望

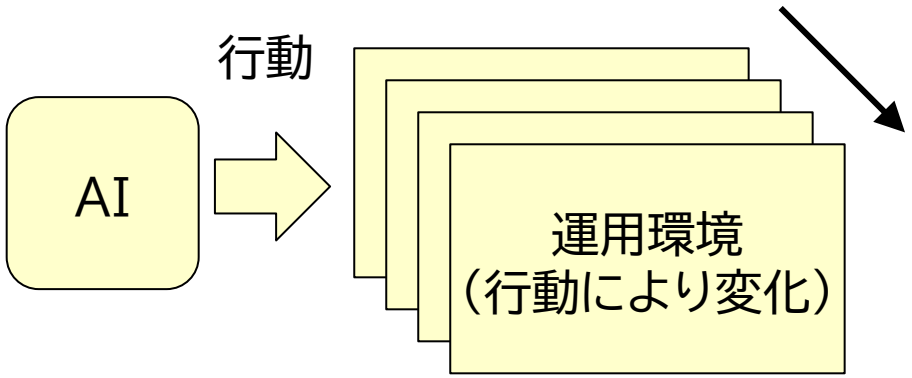
- 大きく異なる複数の運用環境に対応できることが期待されている
  - 店舗、住居、居間、キッチン
- 一つの運用環境においても多様な状況への対応が必要になる
  - ボタン／スイッチ／ペダル／レバー、あるいは、肉／野菜／香辛料
- 運用環境や状況によって、同じ振る舞い（行動）の影響が異なる
  - 水の入ったコップは傾けると水がこぼれる　⇒流し台の上ならOK、PCの上ではNG
- 設計者の期待通り動くとは限らない他者がいる
  - 人（子供、要介護者を含む）、他者が設計したエージェントやロボット



AIの行動範囲の拡大に伴う課題2（未来のAIロボットを例に）

今後の展望

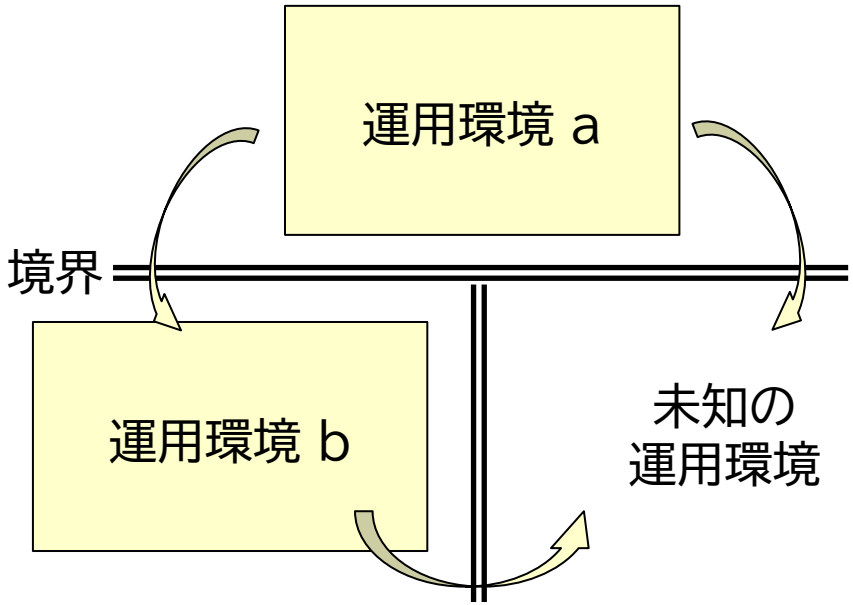
- ある行動を行うことによって、しばらくたってから危害が生じることがある
  - コンロに火をつける、蛇口から水を出す　⇒後で止める必要を作り出している
  - 将棋や囲碁の敗着のように、**後々の行動を制約**して危害の発生につながる行動



AIの行動範囲の拡大に伴う課題3（未来のAIロボットを例に）

今後の展望

- ある運用環境の外に出てしまったことをAIが検知できる必要がある
  - キッチンから居間へ、屋内から屋外へ、敷地内から公共空間へ
- ある運用環境をうまく扱うことと、その**運用環境の境界を検知する**ことは別のタスク
- 別の運用環境に移動したときに、**振る舞いを切り替えられる**か
- 未知の運用環境に入ったときに、慎重な振る舞いができるか



まとめ

- AIの急速な発展によって、品質マネジメントの前提が次々に変化
  - 用途から人が正しい仕様を導ける　⇒ データが仕様を決める
  - モデル構築を自分で行える　⇒　他者が作ったモデルを再利用する
  - モデルの出力はラベルや数値　⇒　モデルの出力は表現物や行動
  - 特定用途向けにモデルを開発する　⇒　汎用基盤モデルに指示する
  - AIの運用環境は設計者が決める　⇒　運用環境は予期しきれない

- AIの技術の成熟を待って規準づくりをするのでは間に合わない
  - 技術の進展の先読みと、それに基づいて先回りした基準作りが必要

- 実世界についてAIに何を教える必要があるかが大きな課題
  - 今はどの運用環境にいるのか、いま行う行動が後で何を引き起こすか

ご参考：AI品質に関するコンソーシアム

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/

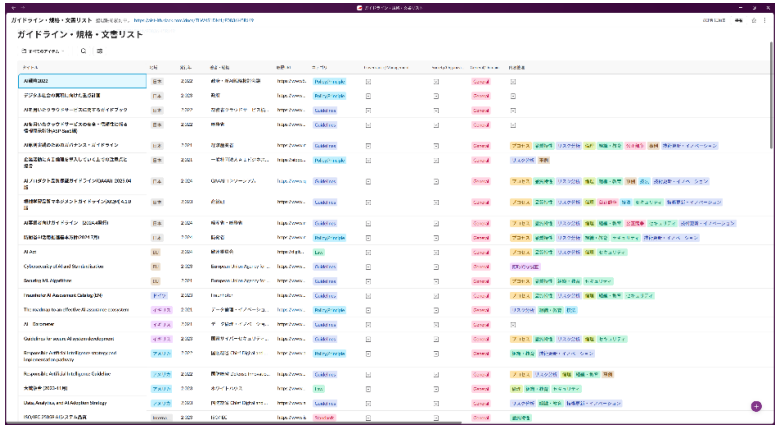
産業界メンバが協調領域で連携してAI品質マネジメントに取り組む

2024年9月活動開始
2027年10月現在 個人会員：86組織から180名以上が参加
法人会員：34社

会長：妹尾義樹（産総研）
副会長：難波孝彰（産総研）
事務局長：小西弘一（産総研）

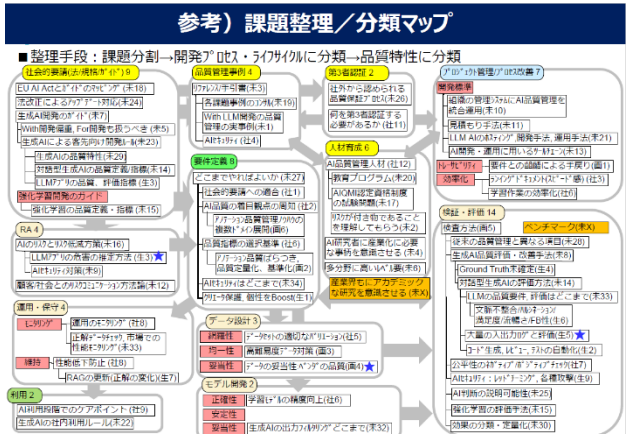
WG1:主査 新原敦介（日立製作所）

情報共有・発信



WG2:主査 三島浩一（DynamicRiskM.）

課題解決



WG3:主査 山田敦（IBM）

エコシステム育成

The End

10th Grand Canvas: AI品質の未来を共に描く ~AI品質マネジメントネットワーキングシンポジウム~

日時：**2027年2月2日(火)** 13:00-17:30 場所：東京駅徒歩圏の貸会議室 ＋オンライン