機械学習品質マネジメント ガイドラインとエコシステムの構築について

2024年6月27日

妹尾　義樹
国立研究開発法人産業技術総合研究所

妹尾 義樹　（国研）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
1986年　NEC入社、中央研究所研究員
1994年～1995年　米国Rice大学客員研究員
1996年　京都大学　博士（工学）
2007年～2011年　NEC北米研究所Director
2018年　産業技術総合研究所、人工知能研究室長
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

主な業務経歴：

学生時代から2005年にかけてスーパーコンピュータ研究開発に従事。専門はスーパーコンピュータアーキテクチャと自動並列化。2002年完成の地球シミュレータは世界最高性能を数年間保ったが、並列処理ソフトウェア開発を統括。2002年IEEE SC2002国際会議においてGordon Bell Award受賞。2007年からBigData, IoTの時代の新たな計算機アーキテクチャ研究を北米研究所にて立ち上げ。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。

機械学習品質マネジメント検討委員会（AIQM）

- 産総研が主催する有識者委員会
  (FY2023までNEDOの委託事業†)
- 「機械学習品質マネジメントガイドライン」を開発・発行してきた
- 委員の御所属 (五十音順)
  - アドソル日進、NEC、慶応義塾大学、
    国立情報学研究所、コニカミノルタ、
    住友電工、テクマトリックス、
    東京理科大学、東芝、トヨタ自動車、
    デンソー、日本IBM、
    パナソニックホールディングス、
    日立製作所、富士通、本田技術研究所

委員長:
中島　震（国立情報学研究所・産総研）

プロジェクトリーダー:
大岩　寛（産総研）

副委員長:
妹尾　義樹（産総研）

† NEDOプロジェクト「人と共に進化する次世代人工知能に関する技術開発事業／実世界で信頼できるAIの評価・管理手法の確立／機械学習システムの品質評価指標・測定テストベッドの研究開発」

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

機械学習ＡＩの品質にまつわる課題

従来のソフトウェア工学の品質管理手法が通用しない。

従来のソフトウェア品質保証
人智が実社会に勝る前提

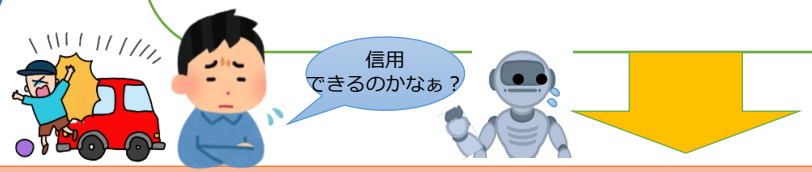
分析的アプローチ
- 危険は**すべて事前に想定できる**と「仮定」

構成的な実装
- トップダウンの品質「作りこみ」
- 問題分析 → 対策の設計 → 実装 → 確認

機械学習AIへの要求：
人の想定を超えるモノの品質保証

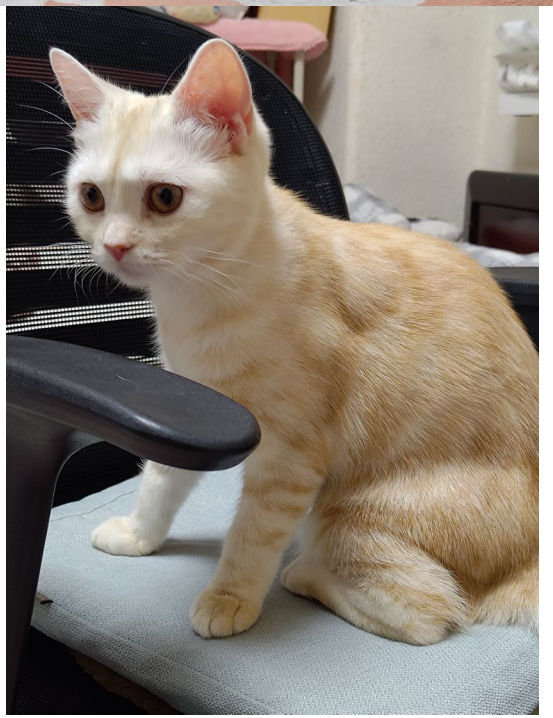
事前に想定しきれない環境への対応
- （例）自動車が衝突する状況を列挙しきれるか？

データから「勝手に生まれる」実装
- ボトムアップ・事例先行型の作りこみ
- 製品の品質の把握も難しい



企業側のジレンマ：「一定の品質」を担保・説明できない

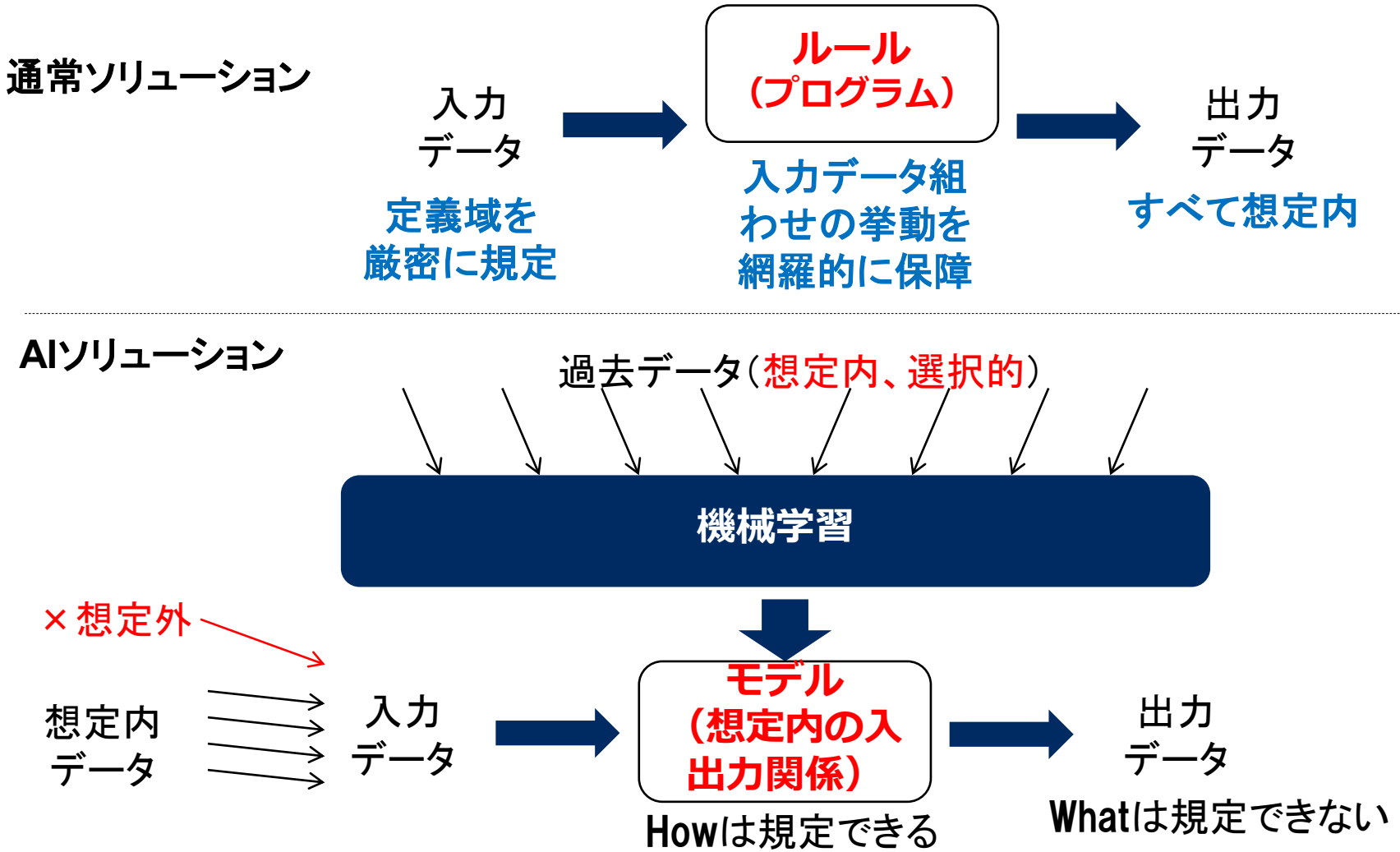
AI品質の難しさ：“猫問題”



例：画像中の猫を検出するプログラム

- 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…）
  - 猫の種類による違い（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…
- 判断基準をプログラムとして記述する方法
  ⇒まず無理、**人には分からない**
- 機械学習による方法
  ⇒検出できるプログラムはできる
  ⇒**判断基準は人には分からないまま**

AIソリューションの機能定義



ビジネスでのAI品質の問題点

AIの品質を担保・説明できず、実績や責任範囲を示せない。

#	AI品質の問題点	望ましい姿
①	顧客に安心して買ってもらえない	•一定の品質を説明できる •説明した品質を実現できる
②	誤動作時にどこまでも責任を追及される	•当初合意した想定の範囲を明確にできる •適切な施策を実施したことを示せる
③	高い品質を差異化要因として示せない	•丁寧に作ったAIの付加価値が示せる

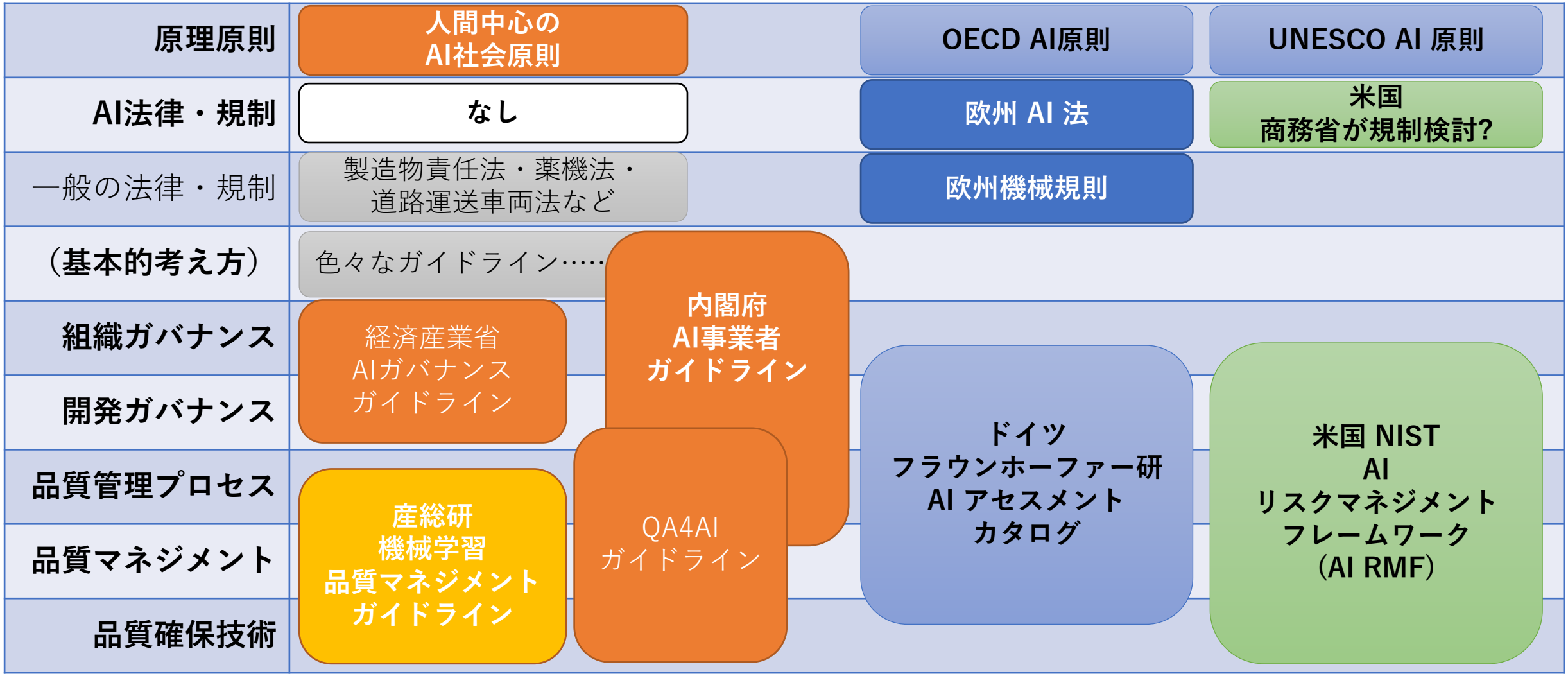
合意形成が必要なポイント

- 要求分析からサポートまで複数の合意が必要

合意形成まで繰り返し
必要に応じて前ステップに戻る

それぞれの固有の合意形成が必要

各国の規制やガイドライン



機械学習AIに求められる品質

品質	左記品質が主に求められる用途の例
有用性	売れ筋予測・リコメンデーション
安全性	自動操縦・運行
公平性	与信審査・人事査定・法執行判断
プライバシー	行動予測、リコメンデーション

倫理

社会的要請

AIセキュリティ　―　上記の品質全てを支える

その他にも色々　―　生成系AIで関心が高まる

機械学習AI品質マネジメントプロジェクト

- NEDOからの受託事業†(2018年度～2023年度)の一環として実施
  - 2024年度からは経産省から直接資金提供を受けて継続中

主な活動

機械学習AIの品質保証手法の体系化

- AI品質マネジメント**ガイドライン**の作成と**国際標準化**
- AI品質マネジメント**リファレンスガイド**の作成

実際に品質を作り込む道具立ての整備

- AI品質管理**テストベッド**の開発
- AI品質**評価向上技術**の開発

† 「人と共に進化する次世代人工知能に関する技術開発事業／実世界で信頼できるAIの評価・管理手法の確立／機械学習システムの品質評価指標・測定テストベッドの研究開発」（2018年度～2023年度）

機械学習品質マネジメントガイドライン

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

URL:機械学習品質マネジメントガイドライン - 成果公開｜デジタルアーキテクチャ研究センター｜産総研 (aist.go.jp) https://www.digiarc.aist.go.jp/publication/aiqm/

ガイドライン：発行実績と経緯

日本語第1版: 2020年6月
リスク回避性とAIパフォーマンス
8つの内部品質

日本語第2版: 2021年7月
公平性の追加・セキュリティ
内部品質 9つに

日本語第3版: 2022年8月
プライバシーの追加
セキュリティの本格化

日本語第4版: 2023年12月
公平性等を含めて体系を一つに整理
内部品質 14個に

英語第1版: 2021年2月

英語第2版: 2022年3月

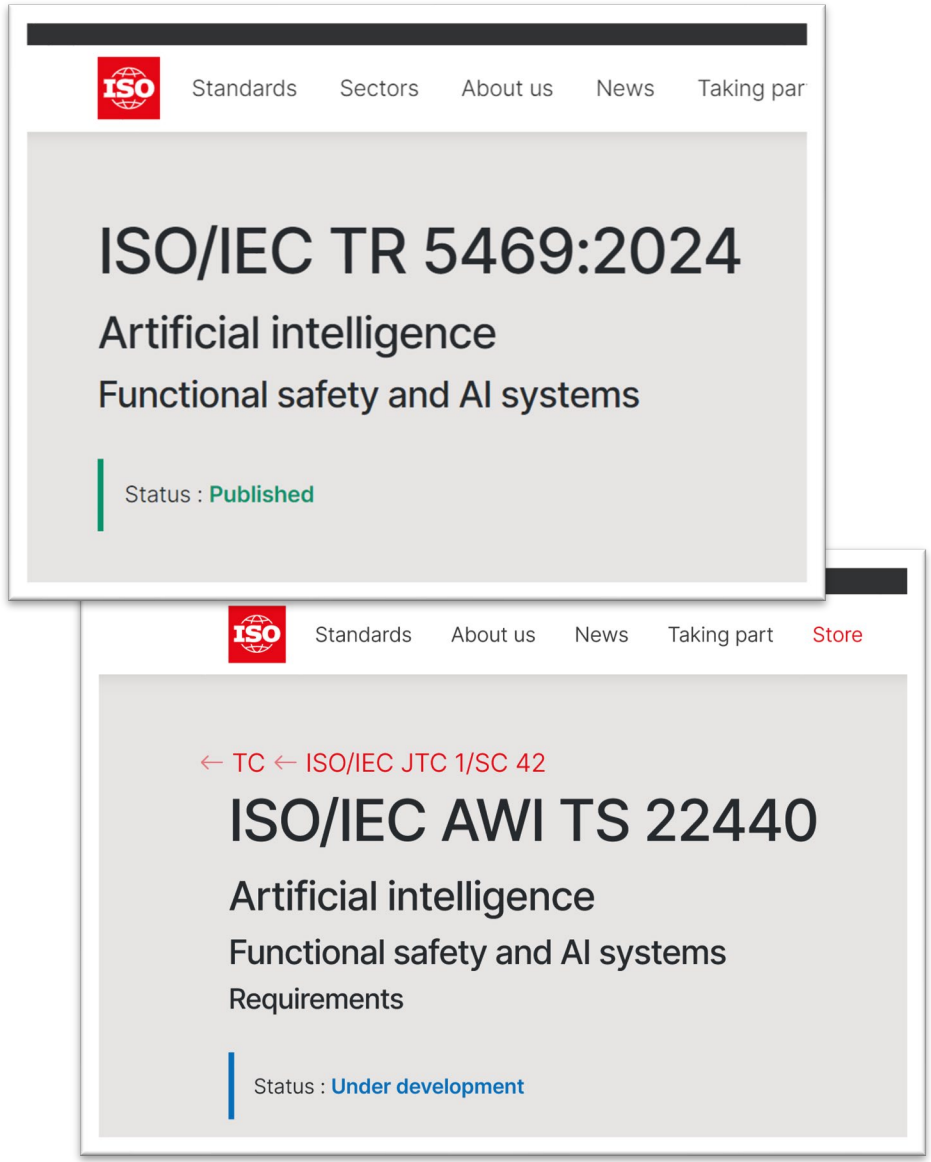
英語第3版: 2023年１月

国際標準化: ISO/IEC TR 5469 (Functional Safety and AI systems)

- 機能安全性 に注目した テクニカルレポート(TR)
- 2024年1月発行

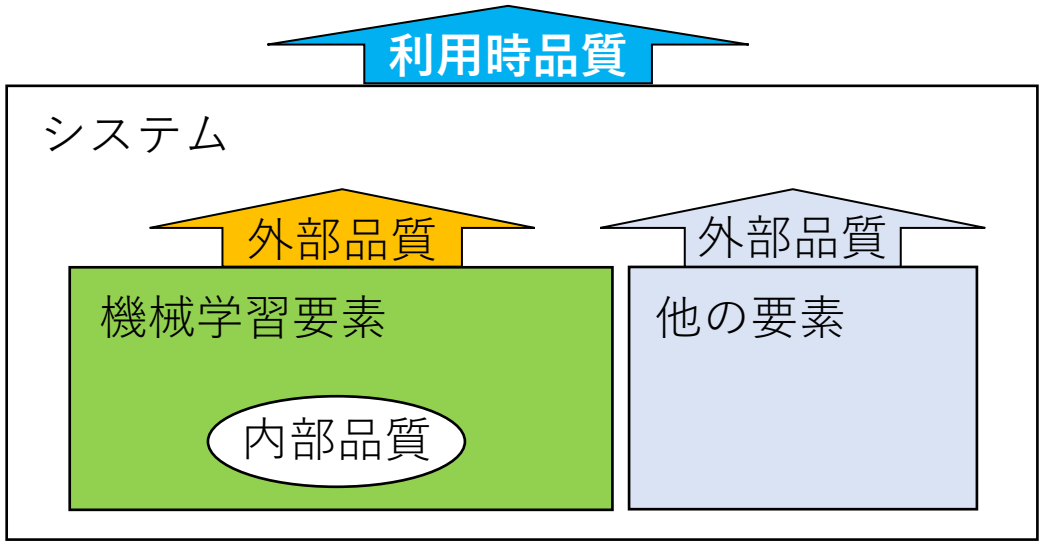
ISO 文書は TR・TS・IS の 3段階

- AIQMガイドラインの内部品質の整理などをインプット
- 担当: ISO/IEC JTC 1/SC 42/WG 3
  - Editor: 日本
- 後続プロジェクト (ISO/IEC **AWI TS 22440**) の検討もスタート

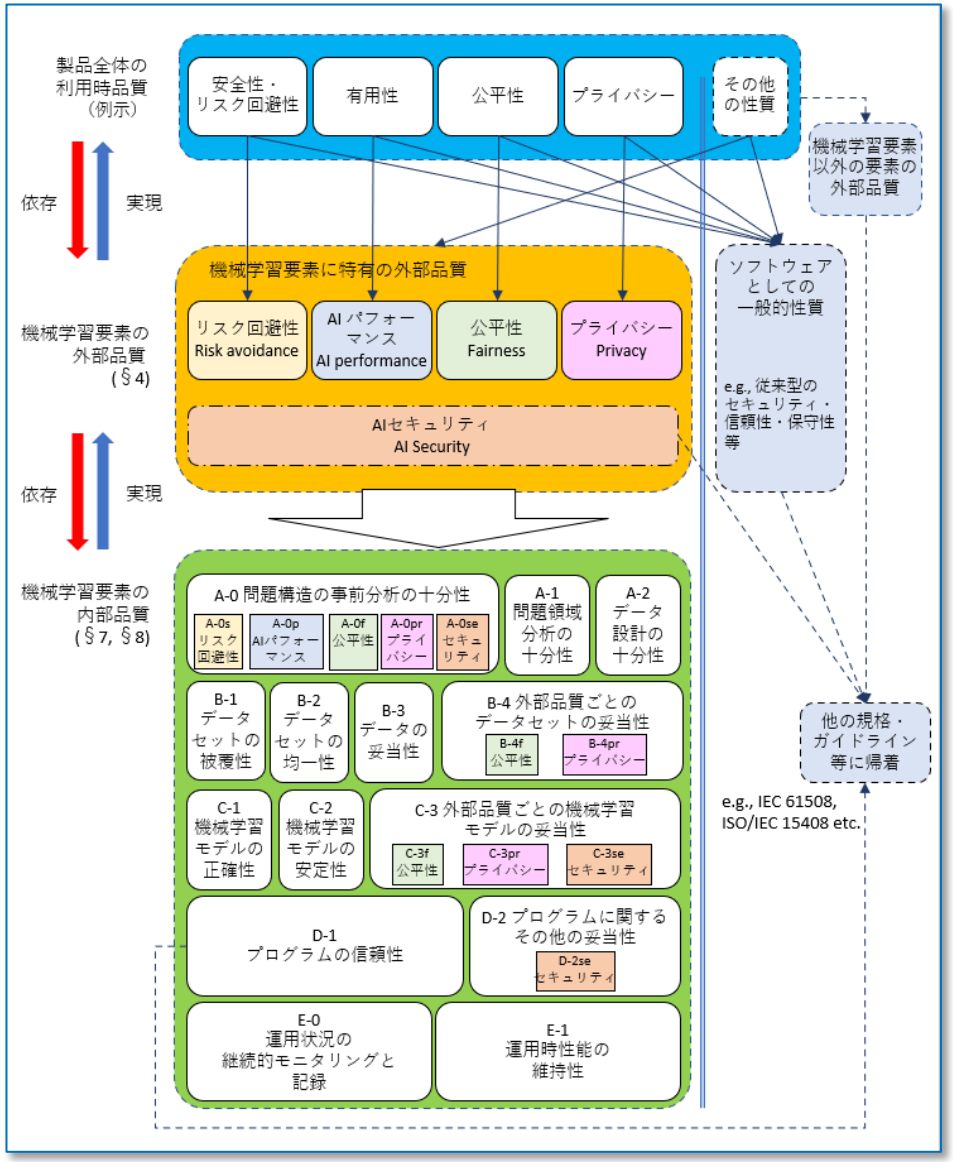


ガイドライン概要

- 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



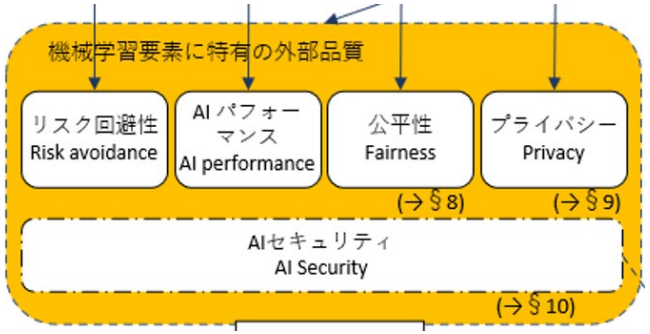
ガイドラインが提案する手法

- 問題認識：機械学習AIでは品質を仕様と動作の整合度として測定できない

- 提案手法：機械学習要素の外部品質と内部品質に着目
  - 機械学習要素が発揮する品質＝**外部品質**
  - **外部品質の向上に資する**、開発中や運用中の観点＝**内部品質**
  - **内部品質の評価改善の達成度**によって**外部品質レベル**を評価

ガイドラインは内部品質を体系的に特定して提示した

外部品質



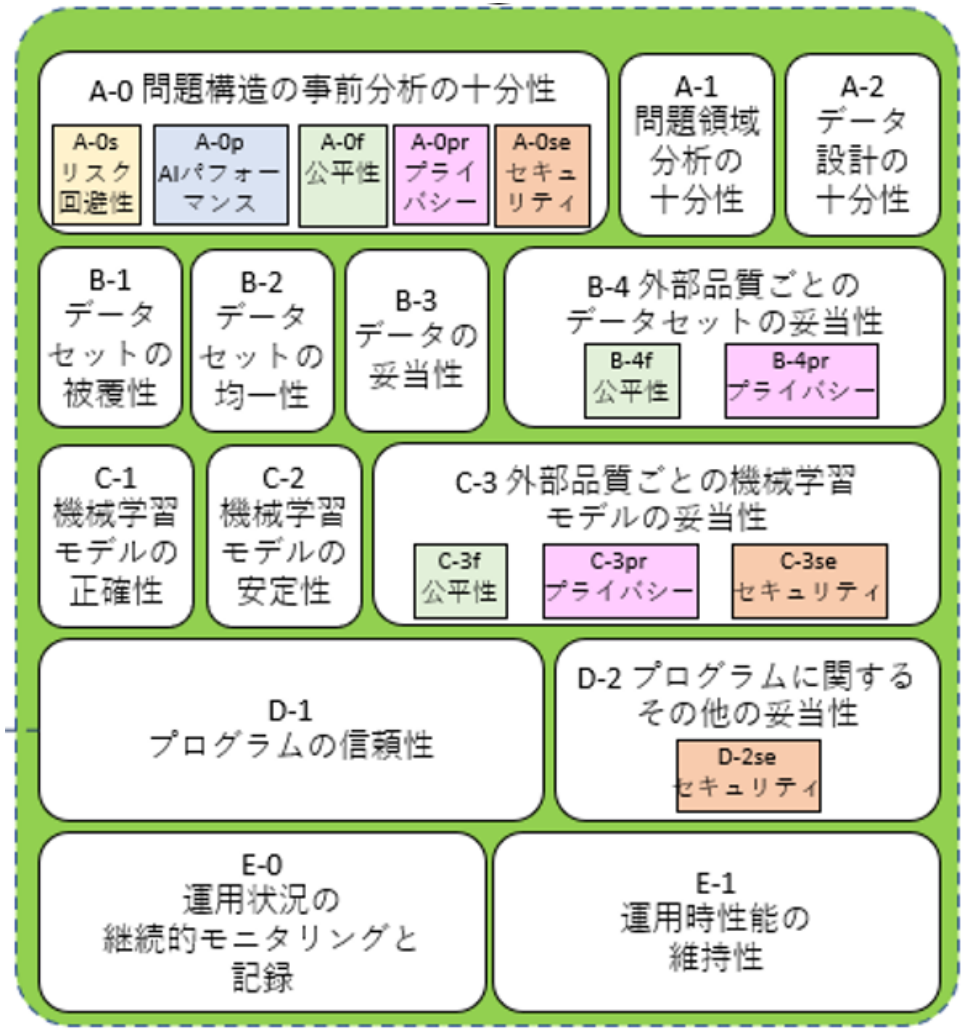
- 品質マネジメントが実現目標とする品質特性

リスク回避性	人的被害や経済損失などを与えるリスクを低減する
AIパフォーマンス	用途に対する有用性
公平性	判断材料とすべき状況以外の要因を判断や結果に影響させない
プライバシー	個人情報や秘密情報の保護を図る
AIセキュリティ	外部要因により他の外部品質が損なわれる可能性に備える

- これらの間にはしばしば**トレードオフ**がある
- 以下は外部品質としては取り上げていない
  - 説明可能性・透明性：品質実現の手段である
  - 倫理性：品質マネジメント以前に決まるものと位置付ける

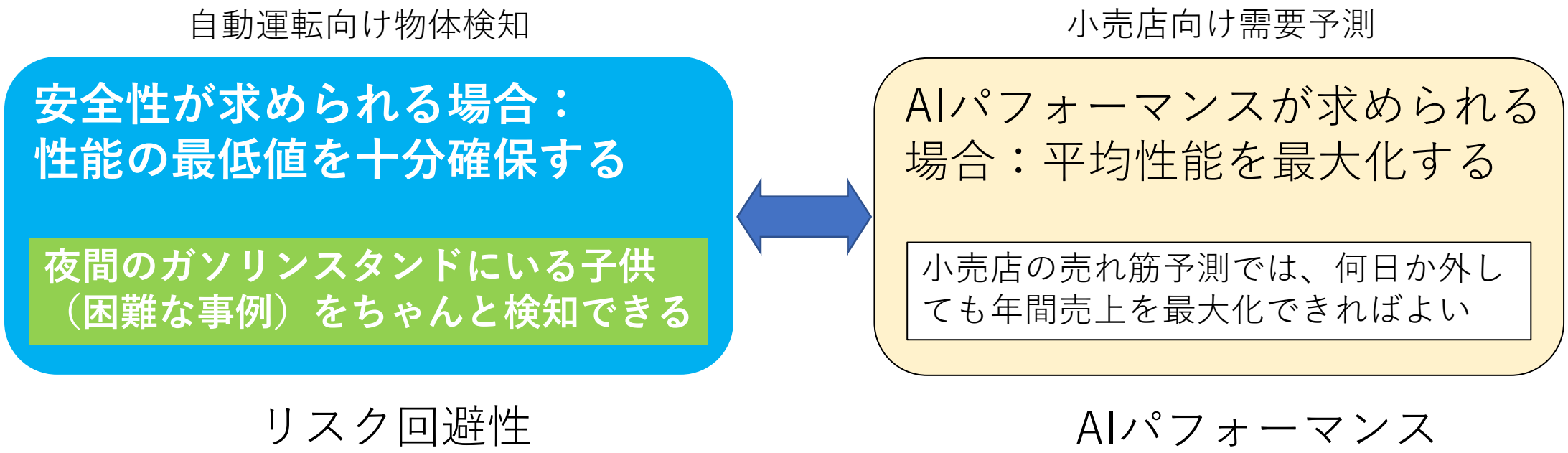
内部品質

- 品質達成に向けて、
  AIシステムの開発運用の中で
  評価・改善の対象となるもの
- 以下の5種類に大別
  - A）**問題領域**の分析とデータ設計
  - B）**データセットとデータ**の妥当性
  - C）**機械学習モデル**の妥当性
  - D）**プログラム**の妥当性
  - E）**運用時**の品質の妥当性
- これらの評価・改善の記録を残すことで、
  品質が外部からも検証できるようになる



例：外部品質間のトレードオフ

- ビジネス要件：自車の周囲の物体を検知する
  - リアルタイム、オフライン、…
- 安全性が重要な品質



例：金融与信審査

- ビジネス要件：クレジットカードの今月の返済が滞りそうな人を検出する
- この事例では公平性も重要
  - 要配慮属性を持つ人に不公平にならないようにする
- 公平性に関して内部品質ごとに行うべきこと

問題分析	データセット	機械学習モデル	プログラム／運用
社会的要請を見極める	データセット内のバイアスを評価し、除去する	モデルの訓練中にバイアスの除去を図る	モデル出力に残るバイアスを検知して除去する

生成AIに関して

- 生成AIや基盤モデルには今のガイドラインは適用できない
- 「従来AI」と異なる想定
  - サプライチェーンがより複雑　⇒全てに責任を持つのが困難
  - 出力の表現力が高い
    ⇒生成物の妥当性評価が難しい（i.e. 記述式問題の採点）
    ⇒新たなリスク（著作権、情報漏えい、他）

- 生成AI向けガイドライン第1版を今年度公表予定
  - 生成AIや基盤モデルを利用してAIシステムを開発する事業者が対象

大規模基盤モデル（LLM）の課題

- 基盤モデル開発者（汎用品開発）とビジネス要件との整合性の担保をどうするか？

どんな情報を提供すれば
品質マネジメントが可能になるか？

生成AIの解釈可能性についての最新情報

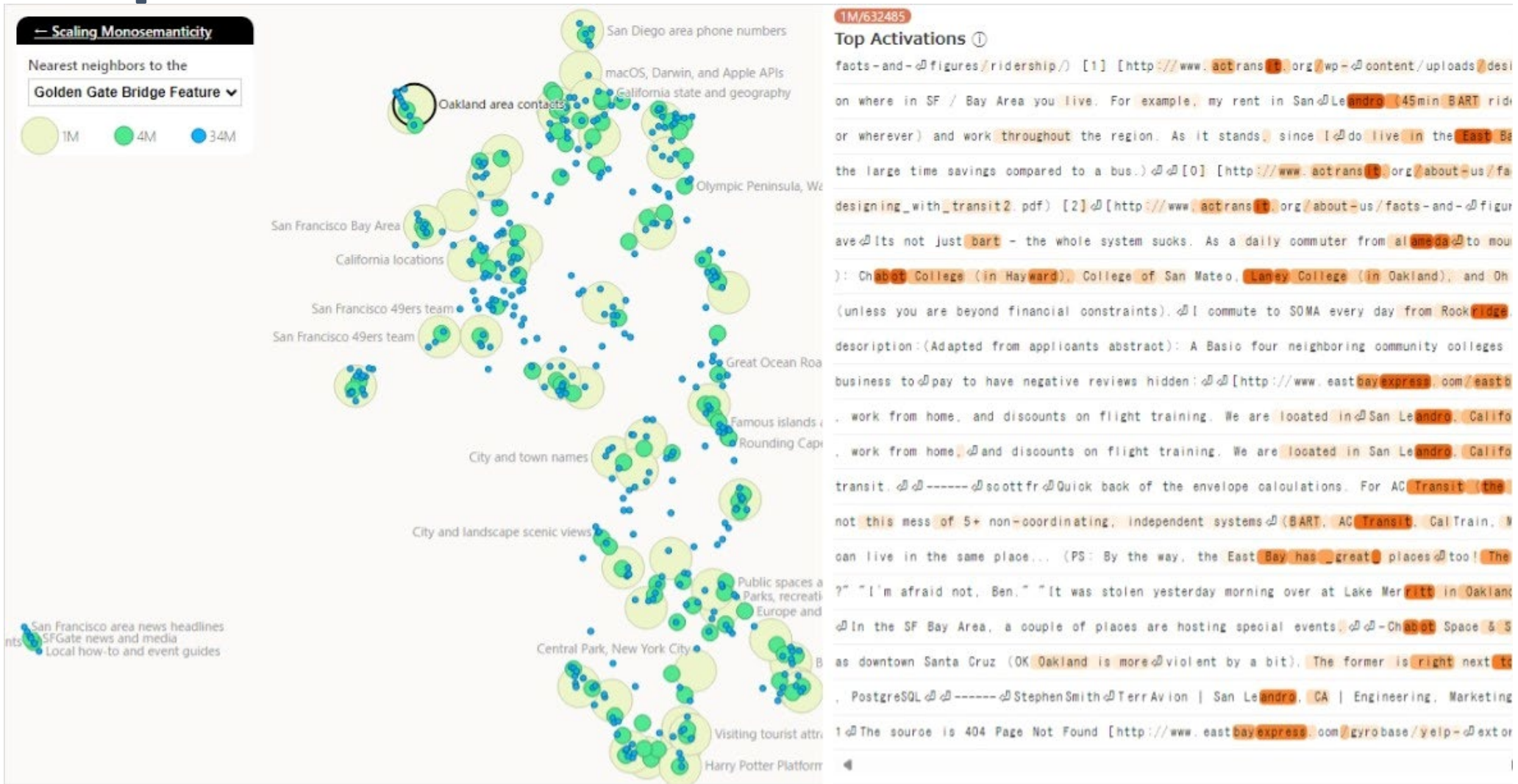
- 実用規模の大規模基盤モデル（LLM）の内部状態から、人が解釈可能な概念に応じたパターン (feature)を数百万個抽出できたと複数社が発表

  - 2024/5/21, Anthropic, Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet
    https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

  - 2024/6/6, Open AI, Extracting Concepts from GPT-4
    https://openai.com/index/extracting-concepts-from-gpt-4/

- LLMの内部状態をSparse AutoencoderによりFeatureの発火パターンに対応づけ
- LLMの思考パターンの解釈ができる可能性

Anthropicのインタラクティブマップ



https://transformer-circuits.pub/2024/scaling-monosemanticity/umap.html?targetId=34m_31164353

これまでの活動と課題、今後の方向

AIQM成果

AIQMガイドライン　リファレンス　テストベッド　品質評価技術　国際標準化

AI品質管理の全体像の明確化、ビジネス活用基盤構築

課題

- 社会的コンセンサス
  - 国際標準
  - 認証
  - 普及促進
- 使いやすさ
  - 手引き
  - 実例集
  - 品質レイヤ整理
- AI評価基盤の構築
  - 実用ツール
  - AIT
  - 教育財整備
- 新たなAI潮流への対応
  - 倫理
  - Security
  - 生成系
  - 基盤モデル
  - ChatGPT

課題解決に向けた活動

企業のおけるビジネス活用

- コンサルビジネス立ち上げ
- 自社内ガイドライン、体制整備
- AIシステム開発への活用
- テストベッド商品化

AIQMイニシアチブ（仮称）

- 企業活動の支援（参加者共助）
  - ✓ベストプラックティスの共有
- 人材育成、教育プログラム
- 宣伝、情報発信
- 活動cへの企業ニーズフィードバック
- エコシステム構築

後継プロジェクト（AIQM２）

- AI品質管理の社会実装基盤作成
- ガイドライン拡充、メンテ
- 国際標準化
- テストベッド拡充、AIT整備
- 実証研究、ベストプラクティス構築
- 品質評価環境・技術整備
- AI新技術対応

AIを安心して使える社会の実現、高品質AIによる日本の産業競争力強化

AIQMイニシアチブ　活動案

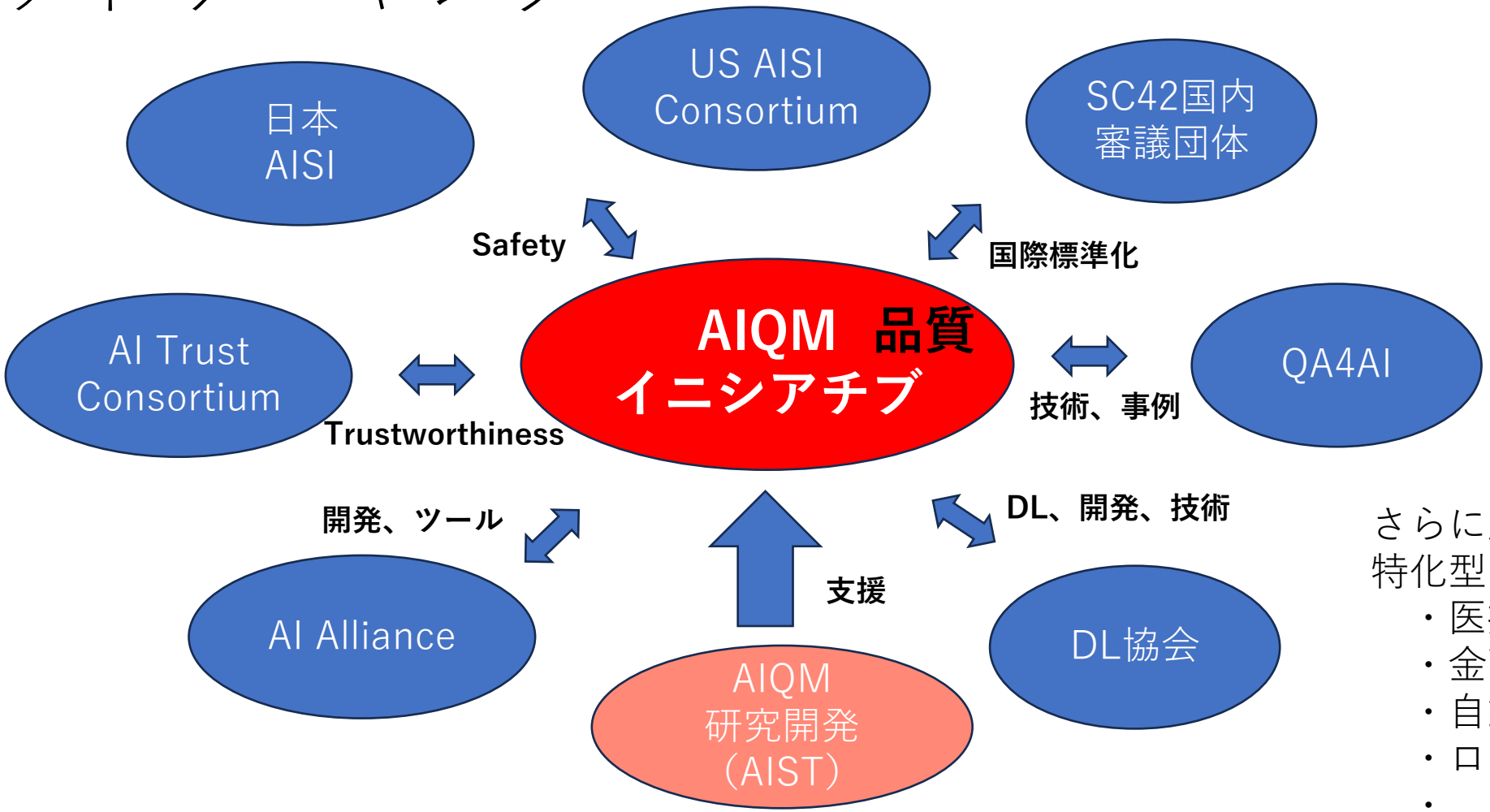
■目的

- AIをビジネスに活用するために必要となる、企業のAI品質マネジメントを支援することにより、AIの安全、安心な活用と、AI活用ビジネスの健全な発展を目指す。

■活動内容

- AI品質マネジメントに関する種々の関連サービスの企画・開発・提供
  - 先端情報取得、手引書、コンサル、ツール、認証
- AI品質マネジメント人材の育成
- AI品質マネジメントに関する社会コンセンサスの醸成のための情報発信
- AI品質マネジメントの国際標準制定についての要望のとりまとめ
- AI品質マネジメントに関するノウハウ蓄積、公開
  - ベストプラクティスの蓄積と共有、先端技術情報の整理と蓄積
- 関連知財の開発、管理、保護
- AI品質マネジメントに関するエコステムの構築支援
- 講演、展覧会、シンポジウム、セミナーの企画、運営
- 書籍、雑誌、印刷物などの企画、制作、販売
- 関連機関との情報交換、連携、協力
- その他、本会の目的達成に必要な事業

AIQMイニシアチブを母体としたネットワーキング



さらに応用ドメイン特化型との連携

- 医療/ヘルスケア
- 金融
- 自動運転
- ロボット
- ・・・

WG1：先端情報の収集と共有、提言

月に2回程度TEAMS会合

- 企業のAI関連ビジネスの遂行に必要な品質に関する先端情報を共有
  - メンバで担当情報収集範囲を決め、これを定期的に共有
- 収集情報（案）
  - 国際標準化
  - 国の施策や方針
  - EUや主要各国の動向（規制など）
  - 技術情報
  - LLMの主要プレイヤーの動向
  - ツール、保険、認証などの状況
- 提言とりまとめ
  - 国際標準化への要望、国への要望、研究開発への要望などを議論してとりまとめ

WG2：ベストプラクティス、手引書

- 現在のガイドラインについて、実用に際しての問題点の洗い出しと解決策の検討
- 問題点の洗い出しを行い、項目分けして解決策を検討
- 項目ごとに解決策を手引書にまとめる（産総研Pjと連携）
- 項目案
  - どこまでやればよいか
  - 外部品質ごとに（Risk Avoidance, AI Performance, Security,,,）
  - 国際基準などへの適合性評価の観点
  - 応用ドメイン（AI活用パターン）ごと
  - LLM対応
  - ツール活用、品質確認作業の効率化の課題
- 無償コンサル機能（問題を持ち込んでもらい、解決策をメンバで検討）

WG3：エコシステム構築

- エコシステムのあるべき姿の検討を行う
  - 協調領域と競争領域の切り分け
  - 一社で難しいことを複数社が集まって成就させることが目的。
- エコシステムそれぞれの機能（ツール、認証、監査、保険など）について、先端情報の調査
- それぞれの機能としてあるべき姿の詳細をまとめる
- 参加企業の強みについて販促の場としても活用
- ツールなどについては、ハンズオンなどのイベント主催
- 情報発信
- 人材育成（教育プログラムの提供）

関連イベント

人工知能標準化国際シンポジウム
～Trustworthy AI実現へ向けたAI標準化～
2024年7月3日～5日
日本科学未来館 7F イノベーションホール
https://www.digiarc.aist.go.jp/event/ai-standardization-symposium-2024/

2nd 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～
2024年7月31日
AP東京八重洲11F Room L & オンライン


産総研