機械学習品質マネジメントガイドラインの発展

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
大岩 寛

社会からみたAIへの恐怖と要求

- AI利用技術の急速な普及
- AIの「得体の知れなさ」「社会的影響」への恐怖が顕在化

⇒ 規制や合意で制約を掛ける動きが起こる

2019年頃：社会原則レベル

- **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)
- **OECD Principles on AI**
  (2019. 5. 22)

など

2020年以降：法制レベル

- **欧州AI法案**
- **日本 AIガバナンスガイドライン**
- **米国 AIリスクマネジメント フレームワーク (NIST)**

次の段階：

技術ガイド
技術標準
など

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:**
  - サービス利用者：サービス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

取り組みの狙い

① 社会全体でのAIの受容性向上・安全性向上
- 劣悪なAIの排除による**利用者**の安全性の向上
- 製造物責任の基準明確化による**提供者**のリスク軽減

② AI構築のサプライチェーンの健全性・競争力強化
- AIのサプライチェーン全体での品質管理
- 受発注基準の明確化によるビジネスの障壁除去
- 製品価値のメトリクス提供による日本産AIの競争力の明確化

⇒ 「安心を説明でき、納得して使えるAI」の実現を目指す

取り組みの狙い

③ 国際的なルール形成とハーモナイゼーション

- 今時のシステムは世界市場を考慮せざるを得ない
- 品質管理はコストの掛かる面倒なプロセス

- もし地域ごとにルールが違うと
  - **地域数だけ作り分けのコストがかかる**
  - **ルールの押しつけと囲い込み合戦**

⇒ 少なくとも方法論は揃えておき、単一プロセスで実施したい

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

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

ガイドラインの位置づけ

ガイドラインの品質確保の構造

詳細は昨年のシンポジウムのビデオをご覧下さい！

品質目標: 5特性 × レベル
（外部品質特性）

- 機械学習AIが「持つべき**品質**」の目標

① **リスク回避性**（安全性）
  - 7レベルを設定

② **AIパフォーマンス**（トータル性能）

③ **公平性**

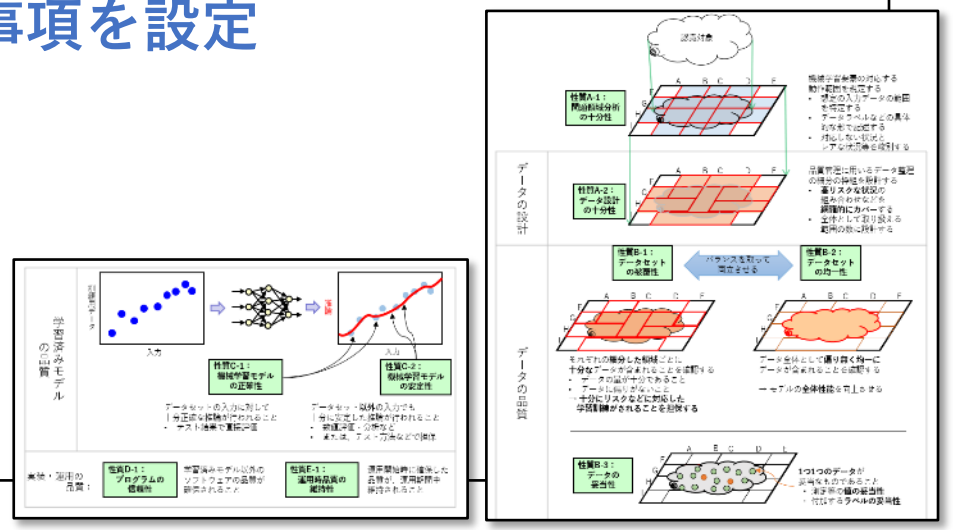
④ **プライバシー**
  - 各3レベルを設定

⑤ **セキュリティ**

品質管理の９ポイント
（内部品質特性）

- 品質を向上させるために押さえるべき技術的ポイントを**９項目に整理**

- **左の品質レベルごとに要求事項を設定**



品質マネジメントの最近・今後の発展

① 国際標準化活動

② AI の倫理性・透明性への要求の増大

③ 基盤モデル・生成AI への対応

④ 品質マネジメント活動の社会展開

① 国際標準化活動

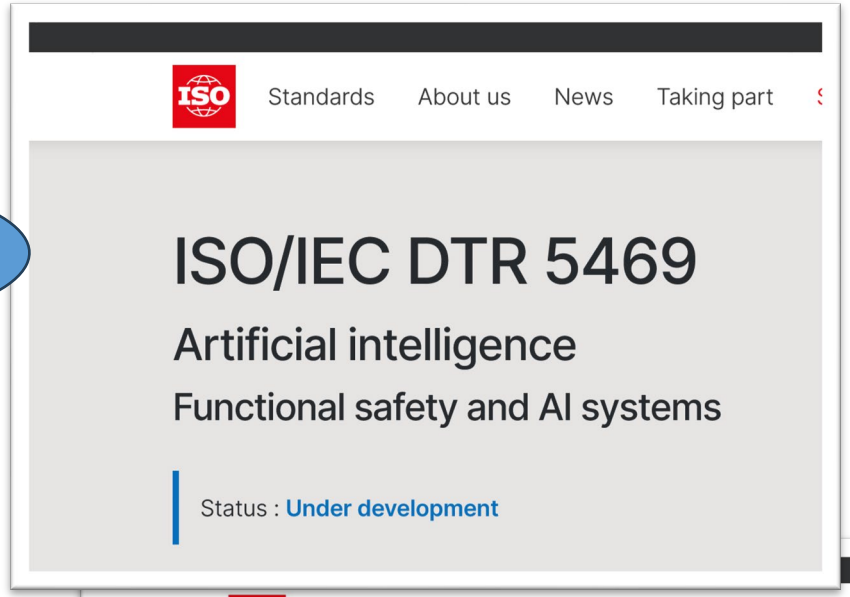
- 国際ルールの統一・協調（ハーモナイズ）を目指すためには、標準化が重要
  - 特に、欧州 AI 法案は ISO 規格や EN 規格の利用が念頭にある
  - 日本のルールを ISO 規格などに整合させること、更には、日本の考え方を ISO 規格に反映することは極めて重要
    - ルールのガラパゴス化は避けないといけない

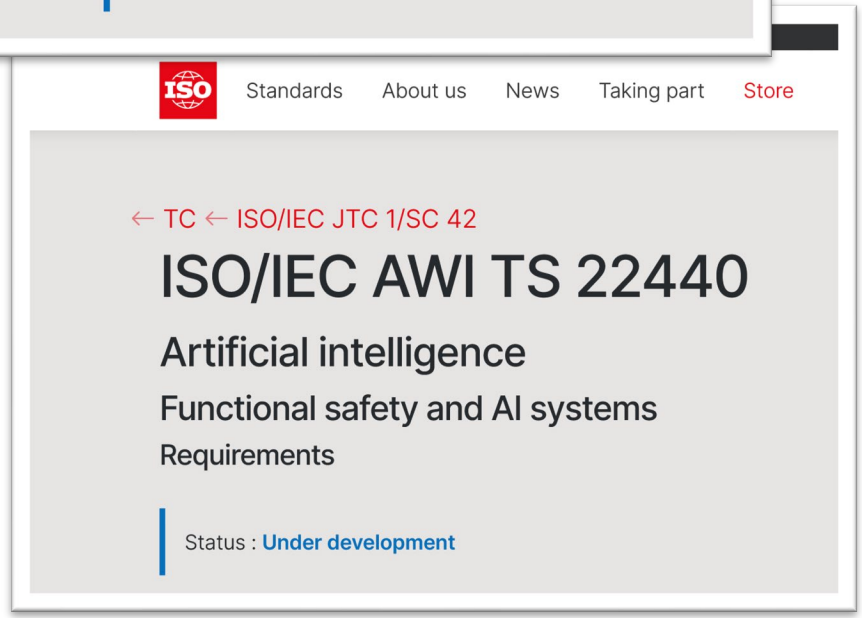
国際標準化: ISO/IEC DTR 5469 (Functional Safety and AI systems)

- 機能安全性 に注目した テクニカルレポート文書

ISO 文書は TR・TS・IS の 3段階

- 担当: ISO/IEC JTC 1/SC 42/WG 3
  - Editor: 日本
  - IEC側 (IEC 61508-3) チームと (事実上合同で) 検討
- AIQMガイドラインの内部品質の整理などをインプット
- 発行へ向けた最終段階 (Stage 50.20)
- 後続プロジェクト (ISO/IEC **AWI TS 22440**) の検討もスタート (Stage 20.00)





② AI の倫理性・透明性への要求の増大

- AI の社会性・倫理性・透明性などへの要求の増大
  - 特に、人を選別する AI に関する関心が強い
  - 欧州 AI 法案でも、この分野の要求は安全性に並んで強い

- ガイドラインは第3版で公平性・プライバシーに関する検討を大幅追加
  - 第4版でも更に拡充・整理の予定

- とはいえ、安全性も引き続き極めて重要と考えています

③ 基盤モデル・生成AIへの対応

- **ChatGPTに代表される「新機軸AI」は AIに対する世相を一変させた**
  - 技術的に品質マネジメントが**一層難しく**なった
  - AIに関する「**得体の知れない怖さ**」が再燃した
- 世界的に協調して取扱い（規制等）を考える流れ
  - G7広島サミット後のプロセスなど
- 機械学習品質マネジメントガイドラインとしても強い興味
  - 第4版～第5版にかけて、新たな対応を考えています

④ 品質マネジメント活動の社会展開

- AI に対する品質実装は、社会実装のフェーズに
  - 一定程度の実践を積んできた
  - 欧州・米国など各国からも類似の考え方が普及してきた
  - また、国際ルール化の流れが ChatGPT 後に大きく変わった

- 機械学習品質マネジメントプロジェクトも次の段階へ
  - 新潮流への対応・既存の安全性などの強化・標準化対応を継続

  ＋ **社会実装の支援などの機能強化を図りたい**

機械学習品質マネジメント特別講座

- 機械学習品質マネジメントガイドラインを中心とした社会人・企業向けの技術講座
  - NEDO の成果普及事業としての支援を頂いています

- 今期パイロットプログラムを開始
- 来期以降、本格展開の予定
  - 並行して、コミュニティ形成のための活動もスタート予定

品質マネジメントの今後の発展へ向けて

- 産総研として日本の産業と、社会の安心・安全のために
  - **協調領域としてのルール形成や標準化**
  - **競争領域としての各社の取り組みの後方支援**

    に取り組んでいきます

- 今日は、実際に民間の実践状況を紹介していただく内容を企画しました

- 今後もご指導のほどお願いします