Rev. 2.5.3.1 (2021-12-08)

機械学習品質マネジメント ガイドライン策定と 標準化の取り組み

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

1

全体の流れ

- 導入・背景
  - なぜ機械学習AIに品質が必要か？
  - 周囲の状況
- 機械学習品質マネジメントガイドライン
  - 検討の経緯・体制
  - 内容
- 今後の取り組み

3

なぜ AI ~~に品質~~が必要か？

- そもそもなぜ「AI」が必要か？
  - 機械学習は現状では品質管理に不利
    - 従来の品質管理手法が通用しない
    - 動作原理がブラックボックスに近い
      - 品質を確保したことをなかなか説明できない
      - なぜ誤動作したかもよく分からない
  - 不安要素を解消できない = 安心できない

…にも 関わらず

4

なぜ AI ~~に品質~~が必要か？

- そもそもなぜ「AI」が必要か？
  - 従来のソフトウェア構築にも限界
    - 現実世界の複雑さに、
      従来のソフトウェア構築手法が追従できない
    - 人間が「正しい方法」を説明しきれない処理は、
      ソフトウェアでは記述できない
      - 「勘と経験」の領域・総合的判断
    - 人ができていないことを計算機にやらせたい

↓

「仕方なく」機械学習を使う

5

産総研 | デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

社会からみた AI への恐怖と要求

- 人間中心のAI社会原則 (2019. 3)
  - 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**
    - いわゆる安全性・信頼性も含まれている
  - 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

6

産総研 | デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

社会からみた AI への恐怖と要求

- 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
    - 何らかの方法でこれらを**担保・実現・説明**する必要がある

7

社会からみた AI への恐怖と要求

- EUのAI規制案
  - 倫理的な側面と安全性的な側面を含む包括的な規制
  - AIをそのリスクでいくつかに分類
    - 最大のリスク分類では使用禁止
    - 高リスクの分類には品質に関する対策を求める

8

AI の品質・セキュリティ面のリスク

- AI システムもITシステムの一種
  - その意味では、従来と何も変わらない
- なぜ、特にAIシステムの品質が難しいか？
  - 使われる環境の**多様性**が大きすぎる
    - 本質的に「**サイバーフィジカルシステム**」のリスク
  - ソフトウェアとしての**構造が特異**
    - 本質的な「**機械学習ソフトウェア**」のリスク

9

産総研 デジタルアーキテクチャ研究センター CPSEC サイバーフィジカルセキュリティ研究センター AIRC 人工知能研究センター

従来のソフトウェア品質の考え方

- 構造的・構築的な品質の作り込み
  - 事前に全てのリスクを**網羅的に列挙**する
  - 各**リスクに対策**をセットし、
    実装の部分に割り当てる
  - 設計に沿って各リスク**対策を個別に実装**する
  - 検査工程で、全ての**対策の実現**を**確認**する

  - **全てのリスクに対策を実現した = 安全** という思想
  - **開発プロセス管理**を通じて抜けがないことを担保

10

産総研 デジタルアーキテクチャ研究センター CPSEC サイバーフィジカルセキュリティ研究センター AIRC 人工知能研究センター

機械学習ソフトウェアの特異性

- 機械学習AIはデータから統計的に構築
  - リスク要因を学習させても、
    常に正しく判断するとは限らない
  - 学習結果の構造が判らないので、
    検査をしても網羅性を確保できない
  - 修正をすると、他の所に未知の影響が出る

11

高品質な機械学習ソフトウェアへ向けて

- 従来のソフトウェア工学の前提が崩れている
  → そのまま技術適用しても結果が保証されない
- 機械学習AIに適合した
  新たな「品質の作り込み」の枠組みが必要

12

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

13

産総研 | デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

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構築の水平分業バリューチェーンの競争力強化
- 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

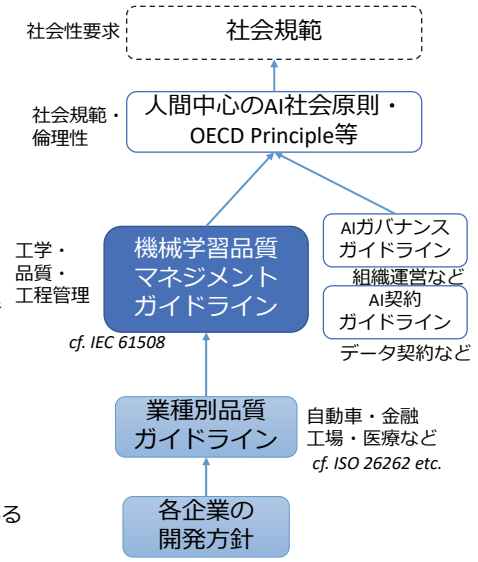
⇒ **「安心を説明でき、納得して使えるAI」の実現を目指す**

14

産総研 | デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

ガイドラインの位置づけ

- 技術的な側面から AIの社会適用を支えるガイドライン
  - 社会規範ガイドライン類の下位
    - 「正しさ」の定義はしない
    - 実現する為に**何が必要か**を整理
  - IEC 61508 などに相当
    - 汎用性を持ったガイドライン
    - 業種特有の具体化は有り得る
- 「自社ガイドライン」を各企業が作るベース
  - **どう実現するか**には任意性がある



15

産総研 | デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

法律・原則等との関係

- 社会性要求（倫理性とか）は法律や原則の類で
  - 拘束的な要求
  - 我々は立ち入らない
- – 我々は社会性を具体的に実現する技術的方法を提示
  - 非拘束的な要求（同じことができれば良い）

16

産総研 | デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

ガイドライン第2版の全体概要

- 文書の構成
  1. 概要サマリー
  2. スコープと定義
  3. 品質レベルの設定
  4. 開発プロセスの参照モデル
  5. ガイドライン適用プロセス
  6. 品質保証のための要求事項
  7. 具体的な技術適用の考え方
  8. 公平性の観点について
  9. セキュリティ対策について
  10. 参考文献
  11. 分析
  12. 図表
  13. 参考文献

207ページ

17

品質管理の対象と考え方

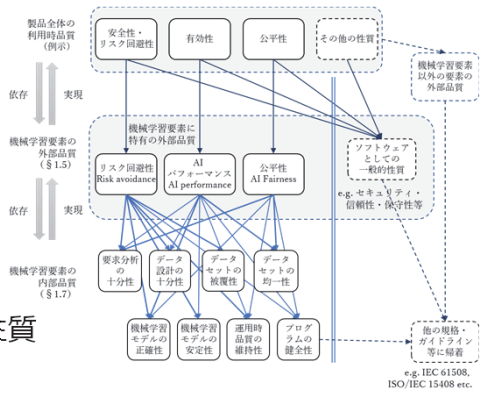
- 「品質」といっても……
  - 立場・視点によって「品質」は違う
  - 利用者にとって:
    - 安心して使える・平等に扱われる・快適 など
  - システム提供者にとって:
    - 危害を加えない・高性能・経済的・公平
    - 目的の要求通りに動く など
  - 開発者にとって:
    - 与えられた仕様通りに動く など

18

品質管理の対象と考え方

- 着目する「品質」
  - **利用時品質**
    - サービス利用者にとっての品質
      - 安心・公平など
  - **外部品質**
    - システムに「**求められる**」性質
    - → **目標を設定**
  - **内部品質**
    - システムが「**持つ**」性質
    - → **達成を確認**

それぞれ**依存・実現**の関係



19

産総研 デジタルアーキテクチャ研究センター CPSEC サイバーフィジカルセキュリティ研究センター AIRC 人工知能研究センター

品質管理の対象と考え方

- 超基本的な流れ
  - 利用者に提供すべき**利用時品質**を考える
  - 製品の設計をして、機械学習要素に必要な**外部品質**を考える
  - 外部品質の**品質レベル**を決定する
  - 外部品質レベルに対応した、**内部品質**の要求事項をガイドラインで確認する
  - それぞれの内部品質の達成を確認する

20

産総研 デジタルアーキテクチャ研究センター CPSEC サイバーフィジカルセキュリティ研究センター AIRC 人工知能研究センター

外部品質: 3項目 × レベル

- 3つの特性に整理

① **リスク回避性**（安全性・危害回避性）
  - ある種のセキュリティ・耐攻撃性なども含む

② **AIパフォーマンス**（トータルの予測精度）

③ **公平性**

21

外部品質① リスク回避性

- 危険に繋がる判断をしない
  （する確率を一定以下に抑える）
  - 安全な判断の範囲での質は問わない
  - ほぼ 安全性 + 金銭的リスク

    議論:「安全性」の語で金銭的リスクを想像するか？
  - 目標レベル (**AISL**) は **7 レベル**
    - 従来の安全性基準 （IEC 61508 SIL）準拠の 4 レベル
    - 従来「SILなし」になる応用を 3 レベルに分割

22

外部品質② AIパフォーマンス

- 全体として平均性能を高く保つ性質
  - ビジネス観点では KPI 実現の成否を問う
  - 最悪ケースの値の低さは問わない
  - 一般的な「予測・推論性能」に近い？
  - 目標レベル (**AIPL**) は **3 レベル**
    - AIPL 2: Mandatory requirements (検収要件など)
    - AIPL 1: Best-effort requirements (一般的)
    - AIPL 0: no requests (PoC 的)

23

外部品質③ 公平性

- 入力の属性に依存して、
  統計的に望まない偏りがないこと
  - 性能・安全より優先されるケースに限る
  - 個別の出力では確認できない性質
  - 目標レベル (**AIFL**) は**3レベル**
    - AIFL 2: Mandatory requirements
    - AIFL 1: Best-effort requirements
    - AIFL 0: no requests (PoC 的)

24

内部品質

- どう「高品質の機械学習」を
  作り込んでいくか？
  - ソフトウェア業界には40年来の歴史の積み上げ
    - ウォーターフォール型設計と段階テスト
    - アジャイル型実装と「テスト・ファースト」
  - **これの意味するところは何か？**

25

内部品質

- ソフトウェア工学の積み上げたもの
  - 「どうソフトウェアが**バグる**か」の経験的知見の積み上げ
    - 事前想定の甘さ
    - 設計ミス
    - 設計を実装に反映する際のミス
    - 実装に引きずられた設計の見落とし
    - etc. etc. …
- 機械学習ではまだ整理が進んでいない

26

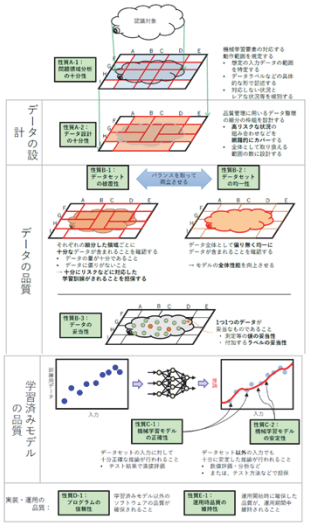
内部品質

- どう整理する？
  - **ボトムアップ・アプローチ**
    - 今AI業界が着目している技術のサーベイ
      - できること、やらなきゃいけないと思っていることが判る
      - それで**十分かどうか**が判らない
  - **トップダウン・アプローチ**
    - AIが「誤判断」した時の原因を仮想的に分析
      - **対処できるかどうか**は判らない
      - 理屈上は、全部潰せば誤判断しない…はず
        - » もちろんそんなことはないけど、改善の方向性としては正しいはず
  - → この2つを併用、特にトップダウンに注力

27

内部品質９項目

A-1. 問題領域分析の十分性
A-2. 問題に対する被覆性
B-1. データセットの網羅性
B-2. データセットの均一性
B-3. データの妥当性
C-1. 機械学習モデルの正確性
C-2. 機械学習モデルの安定性
D-1. プログラムの信頼性
E-1. 運用時品質の維持性



28

内部品質９項目

基本的な考え方

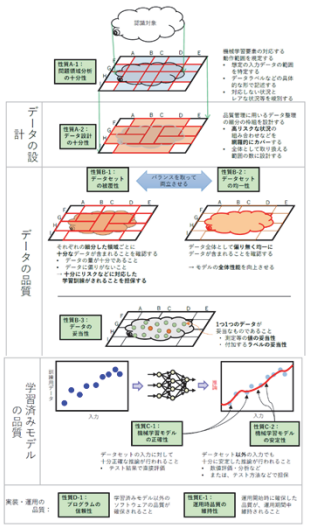
A) 問題の分析に基づく
あるべき**データセットの設計**

B) 設計に合致する
良いデータセットの確保

C) 良いデータセットから得られる
良い機械学習モデル

D) 信頼できる**ソフトウェア**

E) 品質を維持する**運用**



29

データセットの設計に関わる内部品質特性

A-1 問題領域分析の十分性
A-2 問題に対する被覆性
- 品質そのものの「設計」の質を問う特性
- 従来開発の**要求分析・テスト設計**に対応
  - 従来の品質管理でも、実は「ゴールの決定」が大切
- 「**どういう観点で品質を問うか**」を決める
  - 「品質を保つべき状況」を洗い出す
    - 例えば自動運転なら:
      道路の種類・気温・天候・時間帯・障害物の種類 など
  - 複雑な状況の「組み合わせ」を網羅する
    - 例えば「冬の雨」や「夜の一般道」はリスクが高い など

30

データセットの質に関わる内部品質特性

B-1 データセットの網羅性
B-2 データセットの均一性
- A-1〜2で決めた観点に従ってデータ品質を確認
- 基本的には機械学習では
  「十分なデータが均一にある」ことが理想
- だけど…

31

データセットの質に関わる内部品質特性

B-1 データセットの網羅性
B-2 データセットの均一性
– レアケースのジレンマ

- 100万回に1回の頻度の、
  絶対正しく判断してほしい状況にどう対処する？
  - 必ずデータに当該状況が含まれていないといけない
  - 均等に配置すると、レアでないケースが膨大に必要
- 「正解はないけど、必ず考えて設計すること」
  - 敢えて濃度高くレアケースのデータを訓練に使う
  - テスト工程でだけレアケースを追加する など

32

データセットの質に関わる内部品質特性

B-3 データの妥当性
– B-1～2 で集積した個別のデータが妥当かどうか

- テストできる項目と、管理（しかできない）項目
  - データ源の妥当性・信頼性に頼るところもある
  - 検査できるものは出来るだけ検査する

33

産総研 デジタルアーキテクチャ研究センター CPSEC サイバーフィジカルセキュリティ研究センター AIRC 人工知能研究センター

モデルパラメータの質に関わる内部品質特性

C-1 機械学習モデルの正確性
C-2 機械学習モデルの安定性
- データセットから良い「訓練済み機械学習モデル」を作る工程
- 「正確性」「安定性」の2つに整理

技術を社会へ– Integration for Innovation 34 国立研究開発法人 産業技術総合研究所

34

産総研 デジタルアーキテクチャ研究センター CPSEC サイバーフィジカルセキュリティ研究センター AIRC 人工知能研究センター

実装・運用に関わる内部品質特性

D-1 プログラムの健全性
- 基本的には従来の規格に帰着…で良い
  - いくつか特有の注意点があるので敢えて１項目に設定
  - モデル変換とか、エッジデバイス特有の障害とか

技術を社会へ– Integration for Innovation 35 国立研究開発法人 産業技術総合研究所

35

実装・運用に関わる内部品質特性

E-1 運用時品質の維持性

- テストではうまくいっていたモデルが、運用中に品質劣化する可能性
  - **学習訓練時の想定不足**
  - **周辺の状況の変化**

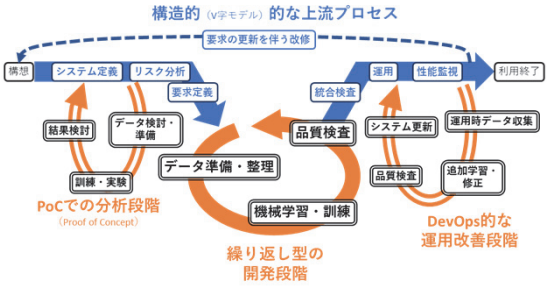
→ これらをモニタリングする考え方が重要

- 一発で品質を作り込むのがなかなか難しい現状に対応
- 追加学習・モデル更新を予め想定して運用を設計する

36

システムライフサイクルプロセス

- 品質マネジメントの全体プロセスモデル
  - 企画段階から運用・利用終了までの総合的な品質マネジメントを想定して整理
  - AI 特有のPoCプロセスや、繰り返し型の開発工程と、品質管理を整合

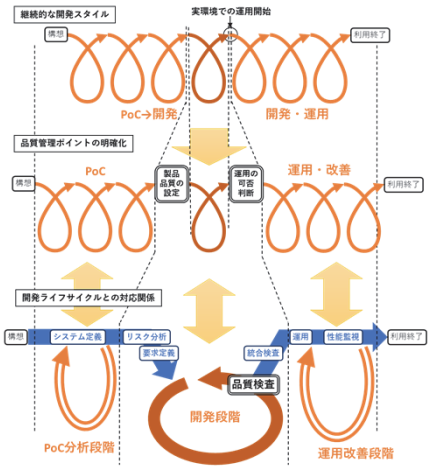


37

産総研 デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

システムライフサイクルプロセス

- 現実のプロセスとの対応
  - **繰り返し型のプロセスでもOK**
  - 品質の**チェックポイント**を明確化させる意図



38

産総研 デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

従来規格との関係性

- 機能安全・SIL (IEC 61508) etc.
  - 安全性が重要な分野では、SIL および分野別詳細規格 (ISO 26262 etc.) への対応はほぼ必須
    - 常識として話が通じる
    - 従来プロセスとの統合が必要
  - AI バックグラウンドの人はほとんど知らない
    - 新たに「SIL 対応しろ」は結構無茶
      - コストがすごく高い

⇒ 両立できるような記述に配慮

39

産総研 | デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

機械学習品質マネジメントガイドライン

- 日本語第1版: 2020年6月公開
- **日本語第2版: 2021年7月公開**
  - 産総研公式ホームページから
- **英語第1版: 2021年2月公開**
  - 英語第2版: 近日公開？
- 逐次アップデート予定
  - 年1～2回程度
  - 実験・研究・リファレンスガイド開発からのフィードバックを元に更新

40

産総研 | デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

国際標準化の流れ (1)

- ISO/IEC JTC 1/SC 42 (Artificial Intelligence)
  - 2019/11 に 東京（産総研）で plenary
  - WG 1 で基本用語集を検討中
  - WG 3 (Trustworthiness) で基本的な議論開始
    - **WG3 に TR5469 “Functional Safety on AI”**
  - その他の品質関連 activities:
    - データ品質 (WG2)
    - AI開発ライフサイクル
    - 品質指標モデル (ISO/IEC 25010 の拡張)

41

国際標準化の流れ (2)

- SC 42 外の状況
  - ISO/IEC TR 29119-11
    Guidelines on the testing of AI-based systems
    - ISO/IEC JTC 1/TC 7
  - IEEE P7003
    - Algorithmic Bias Considerations
  - NIST (アメリカ)
    - AI risk management framework の開発を立ち上げ
    - NIST SP draft, RFI, workshop など

42

TR 5469 (Functional Safety and AI systems)

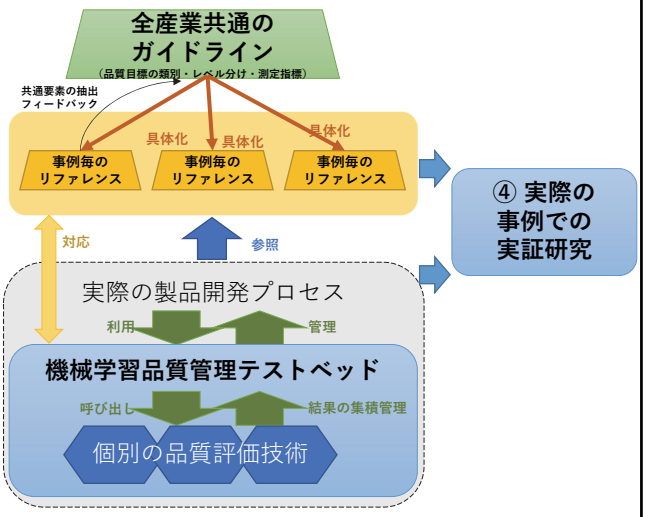
- 担当: SC 42/WG 3
  - Editor: 日本
- IEC側 (IEC 61508-3) チームと (事実上合同で) 検討
- Scope:
  - Use of AI inside a safety related function to realise the functionality
  - Use of non-AI safety related functions to ensure safety for an AI controlled equipment
  - Use of AI systems to design and develop safety related functions.
- 完成目標: 2022/04
- 具体的文言の検討が11月からスタート
  - AIQMガイドラインの内部品質の整理などをインプット

43

その他の活動

①要件の明確化と
エコシステム開発

⇒ 機械学習AI
品質管理ガイドライン
⇒ 産業分野別
AI品質リファレンス



②実際に品質を作り込む
道具立て
⇒品質管理テストベッド
⇒評価ツール

③ 具体的なAI評価技術の
先端開発

45

機械学習品質評価共通基盤（テストベッド）の開発

- **容易に開発できる高品質AIを実現するための環境整備**
  - 高品質AI開発ライフサイクルのための開発と運用を組み合わせたDevOps機能の強化
  - データ準備・機械学習システムの開発から品質検査・評価までのプロセス全体を管理するツール
  - 既存開発フレームワークなどとの関係も検討

46

産総研 | デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

応用別のリファレンスガイド

- **ガイドラインを実際に製品に適用するための事例ベースの「ガイダンス」の作成**
  - 目的1: 汎用で抽象的なガイドラインを補完する、**「こうすればできる」**の提示
  - 目的2: ガイドラインの**実用性の確認**
  - 目的3: 具体的事例からの**フィードバック**
- 民間企業の具体的事例をベースに事例抽出
- 「共有知」を目指した取り組み
  - 成果公開を前提とした研究活動
    - 民間企業から出向して頂き、産総研の立場として研究

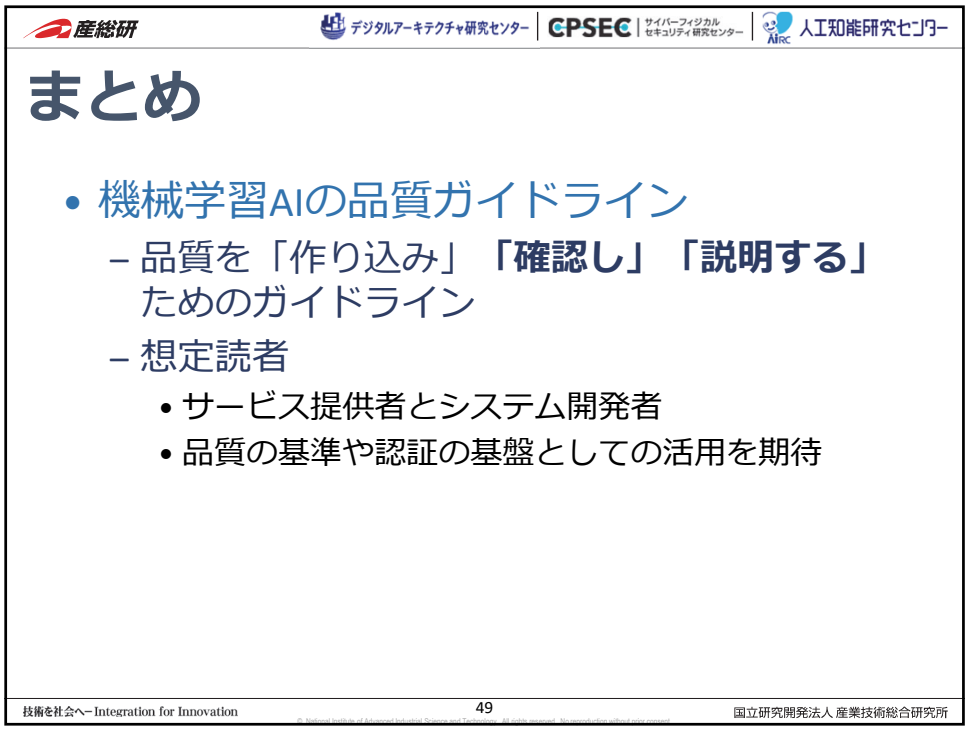
47

産総研 | デジタルアーキテクチャ研究センター | CPSEC サイバーフィジカルセキュリティ研究センター | AIRC 人工知能研究センター

具体的な品質管理技術の研究開発

- QA/QC のための具体的な技術の追求
  - **ソフトウェア工学**の品質管理技術の機械学習への適用
  - 「品質」という指標に着目した新しい**機械学習技術**の研究
  - ガイドライン中で必要とされる具体的な品質指標に対する測定技術の開発

48



49