AIST
Create the Future, Collaborate Together

AISI Japan AI Safety Institute Partnership

機密性1情報　投影用

AIセーフティに関する取り組み

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 副研究部門長
AI セーフティ・AISI パートナーシップ担当

2026 年 5 月 21 日

産総研の取り組み

- 2015年頃からAIの品質に着目
- 2018年に「機械学習品質マネジメント検討委員会」設立
  - 国内の10社以上の品質管理・AI開発者
  - ほぼ週1の議論を今まで継続

- 2020年に「機械学習品質マネジメントガイドライン」第1版を公開
  - 以降、年1回程度の更新と英語版の公開

- 2025年に「生成AI品質マネジメントガイドライン」を公開

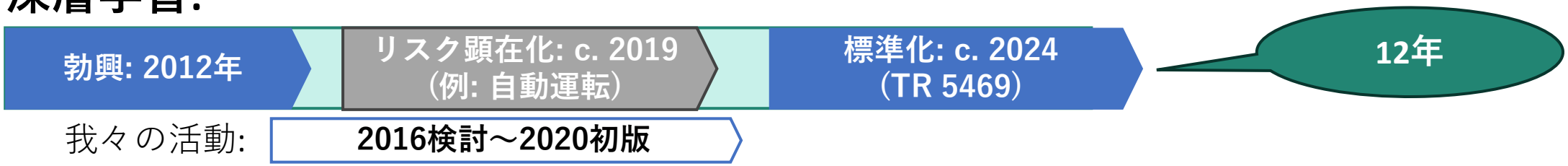
ChatGPT ショックとその後の怒濤

- **2022年のChatGPTの登場**
  - **AI が「得体の知れないもの」に戻った**
- **AI の役割がどんどん拡大していく**
  - 技術的に面白いことはどんどん増えていく
  - 興味ベースの探求は止まらない
  - **実世界AIの拡大**
  - **マルチモーダルAI**
  - **AI エージェント（人の代わりをするAI）**

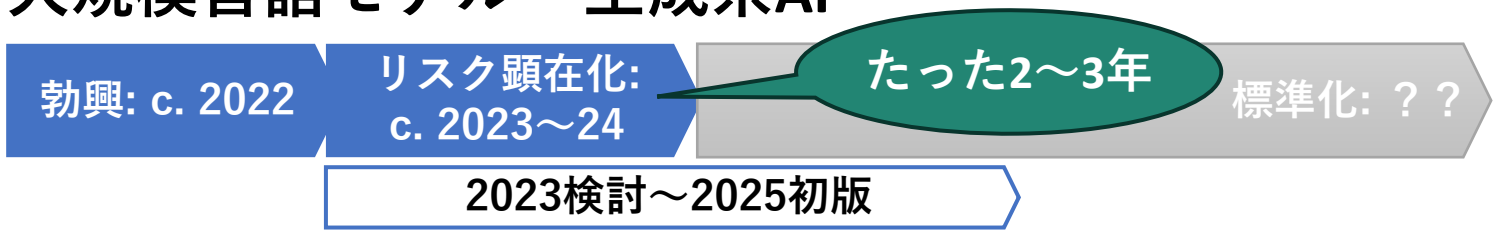
更に迫る危機: 技術進展のとんでもない加速

- 技術の進展 ＝ 安全にとってのリスク ＝ 新しい取り組みの端緒
- 今あるリスク: **技術進展が早すぎる**
  - 従来ソフトウェア

  - 深層学習:



  - 大規模言語モデル・生成系AI



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
- 先を見据えた取り組みと、今日の問題を解く取り組みを両方やる

  - 中長期の取り組み: エージェントの「次」までに何が起こるか？

  - 短期の取り組み: しっかり マルチモーダルAI やエージェンティックAI の
    技術そのものを安全にする

  - 直近の取り組み: 応用ごとに企業と一緒になって、
    ベンチマーク・評価環境など、明日のビジネス展開を安全にする

中期的AIリスク展望 2026

- 今後発生する新たなAIリスクに対応するための「予測展望」
  - 5〜10年後までのAIリスクを予想・整理・構造化し、それに応じた研究課題を示す

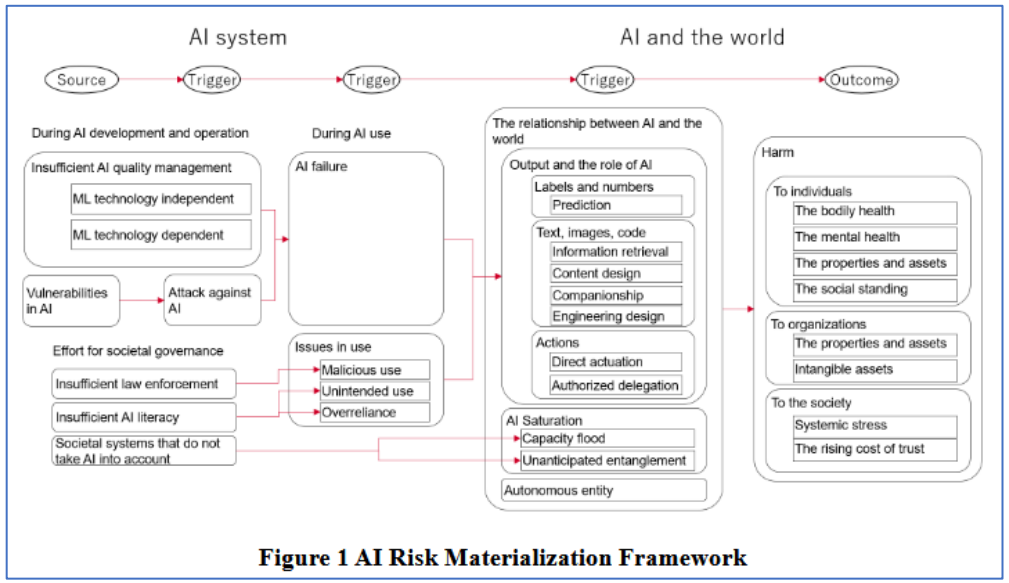


Figure 1 AI Risk Materialization Framework

中長期リスクの具体化フレームワーク

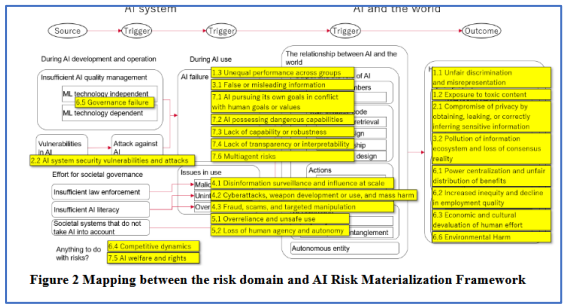


Figure 2 Mapping between the risk domain and AI Risk Materialization Framework

MITフレームワークとの相互比較

(1) Formalization and Articulation of Unstated Assumptions
(2) Machine-Readable Restructuring of Normative Documents
(3) Analysis of AI's Ways of Understanding of Safety and Ethics
(4) Addressing Insufficient Understanding of Real-World Phenomena
(5) Generalization to Overcome Limits of Scenario Enumeration
(6) Connecting Case-Based Learning with Principle-Based Understanding
(7) Overcoming Limitations of Domain-Specific Safety Measures
(8) Detection and Suppression of Out-of-Scope Behavior
(9) Self-Understanding and Judgment Deferral
(10) Systematization of Regional and Cultural Differences
(11) Detection of Contextual Misalignment and Adjustment of Safety Behavior
(12) Judgment Principles for Handling Cultural and Contextual Differences
(13) Integration of Technical, Organizational, and Societal Measures
(14) Robustness Against Emergence of New Architectures
(15) Designing Research and Investment for Discontinuous Technological Shifts

研究アジェンダ15項目

今後・国内外の関係者と議論し更新の予定
（例）国内ワークショップ開催
英ケンブリッジAIリスク研・エジンバラ 未来リスク研
米メリディアン研究所 など

中期的AIリスク展望 2026

- **いくつかのキーポイント**
  - **現実社会（物理・ルール・常識など）理解の解像度の向上**
    - ルール・法律などの機械可読化
    - AI の「常識」への理解についての研究
    - 文化や社会の地域差などへの適応
  - **技術的・法的・社会的対策の整合**
    - 悪意と事故への対策の違いの認識と調和
  - **不連続な技術変化への適応**
    - 多数の（別々の）エージェントが
      実世界側で相互干渉する世界への対応
    - 「Transformer の次」への準備

(1) Formalization and Articulation of Unstated Assumptions
(2) Machine-Readable Restructuring of Normative Documents
(3) Analysis of AI's Ways of Understanding of Safety and Ethics
(4) Addressing Insufficient Understanding of Real-World Phenomena
(5) Generalization to Overcome Limits of Scenario Enumeration
(6) Connecting Case-Based Learning with Principle-Based Understanding
(7) Overcoming Limitations of Domain-Specific Safety Measures
(8) Detection and Suppression of Out-of-Scope Behavior
(9) Self-Understanding and Judgment Deferral
(10) Systematization of Regional and Cultural Differences
(11) Detection of Contextual Misalignment and Adjustment of Safety Behavior
(12) Judgment Principles for Handling Cultural and Contextual Differences
(13) Integration of Technical, Organizational, and Societal Measures
(14) Robustness Against Emergence of New Architectures
(15) Designing Research and Investment for Discontinuous Technological Shifts

研究アジェンダ15項目

マルチモーダルAIセーフティガイドライン

- **マルチモーダルAIに対応した品質マネジメントガイドライン**
  - **対象: テキストと画像など複数のデータ形式を扱う生成AIシステム**
  - **従来ガイドラインを踏襲しつつ、MMAI特有の差異を明らかに**
    - 新しい品質指標:
      異なるモダリティの間での事物の対応関係を正しく把握する能力（クロスモーダル照応能力）
    - AI に期待する能力をレベル分けとして類型化・提示。
  - **３つのマルチモーダルAI応用事例を取り上げ、具体的な検討例も提示**

動向調査や環境整備

- **国際AI規制関連文書マップ**
  - **日米欧＋独（Fraunhofer）のAI品質・セーフティ関係の文書群の比較体系化**
    - 従来AIと、生成AI (Generative AI, General-Purpose AI) 以降についての状況も整理
- **国際AI標準文書マップ**
  - **ISO/IEC, IEEE, CEN/CENELEC のAI安全関連文書類の包括的マッピング・構造整理**
- **生成 AI 実践ガイドと企業事例集**
  - **民間企業による生成 AI の活用の取り組みの調査・体系化**
- **生成AI安全性評価プロトコルとその実装ガイド**
  - **ISO/IEC 42001（AI全般のマネジメントシステム）を具体的に生成AIに適用する考え方の整理と実践**

マルチモーダルAI・生成AI向けの品質管理技術

- 成果の一例
  - ユーザー意図を正確に解釈し不確実な事象にも適切に応答できるマルチモーダルAI
  - 学習データに起因する偽誤情報の防止のための選択的忘却技術
  - 大規模空間データにおける出力信頼度の評価技術
  - 規則に基づくLLMアラインメント技術の安全性評価

応用ドメインでのマルチモーダルAIの安全性実現

- **対象領域やアウトプットの一例**

日常空間＋事故防止	生活空間＋多様性	音声データ＋公平性	公共空間＋安全性

この4つは前半で紹介

科学実験＋安全性

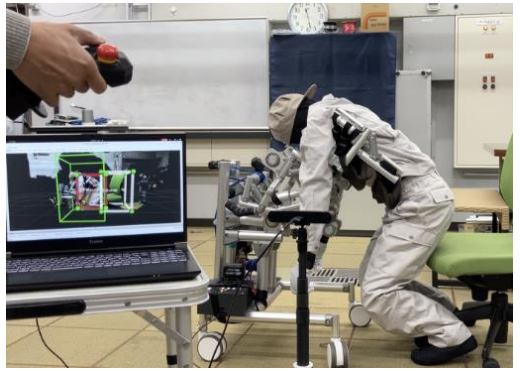
- 科学における法則や知識との整合性がないAI予測結果・生成データのリスクの評価
- 実験の安全性を担保するためのLLM制御手法等の開発

生命科学＋安全性

- ライフサイエンス分野の科学AI向け有害性分子データベース
- 生成アウトプットの有害性予測のプロトタイプモデル

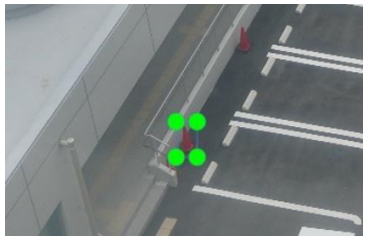
機能安全×AI

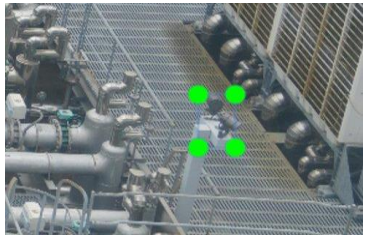
- 人体へ直接接触するロボット制御への多層化安全技術
- 人間拡張AIシステムのセーフティ評価技術



公共空間＋プライバシー

- 人検出モデルの性能向上のための背景学習
- 人と誤認識される背景画像のデータセット

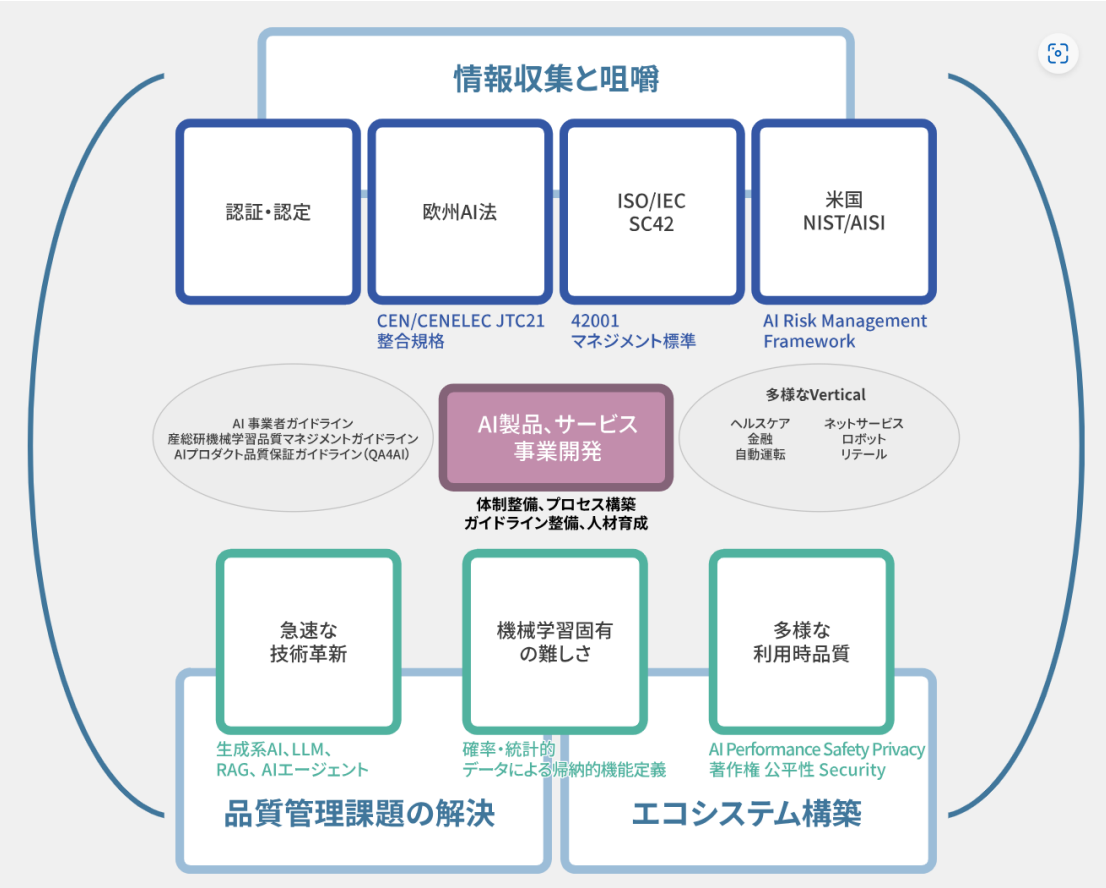




社会への成果普及: AI品質マネジメントイニシアティブ

- **民間企業の相互連携を加速・支援するコンソーシアム**
  - **海外・国内の最新情報収集・咀嚼**
    - 海外法などへの対応議論
    - 国際標準化動向などへの対応
  - **開発現場の知見の蓄積交換・ベストプラクティスの共有**
    - 現場向けの手引き書
    - 開発者向け講座など
  - **開発・テスト・認証などの企業を跨いだエコシステムの構築**

AI品質マネジメントイニシアティブ
AI Quality Management Initiative



AI品質マネジメント講座

- **機械学習AIを利用するシステムの実務者の方々を対象とした社会人向け講座**
  - どう目標設定するか（企画部門、コンプライアンス推進部門の方々）
  - どう要求するか（発注者の方々）
  - どう実現するか（開発者の方々）
  - どう検証するか（品質保証部門の方々）
  - **様々な立場で機械学習AIの品質に取組む方々とのネットワーキングの機会に**
- **開催形式**
  - **1期を全4回で構成、月に1、2回ずつ開催し、2-3か月で終了**
  - **各回とも1日で90分のセッションを3コマ実施**
  - **講師による講義と、少人数でのグループワークによる演習**
  - **対面で開催、受講者20名強**

Create the Future, Collaborate Together