AIST
Create the Future, Collaborate Together

先端AIの品質マネジメントについての今後の展望

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

2025 年 7 月 8 日（5th Grand Canvas）

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

イントロダクション：AIと技術、あるいは社会哲学

- AI黎明期 ～ SF と恐怖の存在
  - HAL 9000 (?) / ターミネーター
    （サイボーグだけど「異質な存在」として）
- 2005年
  - 「シンギュラリティー」
- 2010年代後半
  - AI法規制・AI自主規制の議論
- 2020～2021年
  - AI を技術的に御す方向性
  - **哲学から工学技術へ**

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

社会からみた AI への恐怖と要求: 前史

- **「人工知能」に対する感覚の相違**
  - **日本**
    - 多神教的・擬人化（人のようなもの）に比較的寛容
    - ドラえもん・鉄腕アトム など、敵にも味方にもなる存在
  - **欧米**
    - 感覚的には人と機械を峻別する傾向
    - HAL 9000 （2001年宇宙の旅）など、古くから恐怖の対象として描かれる
- **シンギュラリティー論（2005年）**
  - **AI の「自己進化」に対する恐怖感**

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

社会からみた AI への恐怖と要求

- **2010年代後半**
  - **政治・統治のレベルに議論が（ようやく）到達**
    - 社会として認知して制御するレベルに至る
  - **要求を整理・明示して文書化する動き**
    - 社会原則・Principles レベル
    - 法律・社会的ガイドライン レベル
    - 技術ガイドライン・技術標準レベル

- **2019～2021年～……**
  - **技術としてAIを具体的に制御する流れ**
    - やっと、「得体の知れたもの」として扱うように

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

社会からみた AI への恐怖と要求

- 2019年頃から: AIに対する社会からの要求の明文化の動き

- 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)
  - 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
  - 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

- 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

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

社会からみた AI への恐怖と要求

- **2020年代: 法律・ガイド層の取り組み**

米国

NIST
AIリスクマネジメント
フレームワーク
(2021.7作成開始)

米国政府調達要件に入ると
サプライチェーンに連なる
日本企業にも影響する恐れ

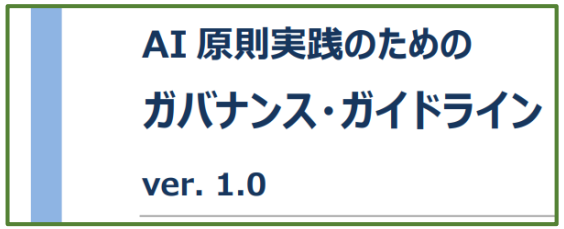
欧州

EU AI act
2021.4.21 法案公表
▶ 2024.8 **発効**

高リスクAI応用を特定
施行されると欧州市場向け
ビジネスで必須に

日本

(2021.7.9) 経産省
AIガバナンス・ガイドライン
(2024.4.19)
AI 事業者ガイドライン
▼
(2025.2) AI基本法案
技術革新とリスク対応のバランスを狙う



(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

社会からみた AI への恐怖と要求

- **日本での流れ**
  - **2019.3 人間中心のAI社会原則** (統合イノベーション戦略推進会議)
  - **2021.7 AIガバナンス・ガイドライン** (経済産業省)
  - **2024.4 AI 事業者ガイドライン** (経済産業省・総務省)
  - **2025.6 AI推進法**

  - **2019.5 AIプロダクト品質保証ガイドライン (QA4AI コンソーシアム)**
  - **2020.6 AI品質マネジメントガイドライン (産総研)**

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

欧州の流れ

- **2019.8 Ethics Guidelines for Trustworthy Artificial Intelligence**
  - **(High-Level Expert Group on AI)**

- **2021.4 EU AI Act 法案提出 ▶ 2024 施行**
  - **ポイントは「域外適用」**
    - 海外の欧州向け・欧州人向けサービスも規制対象になる
    - 全世界の売り上げに対する割合での課徴金

- **現在 技術レベルの規則となる「整合標準」を策定中**
  - **CEN/CENELEC**

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

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

謝辞: NEDO委託研究「**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**」（JPNP20006）の成果を含みます。

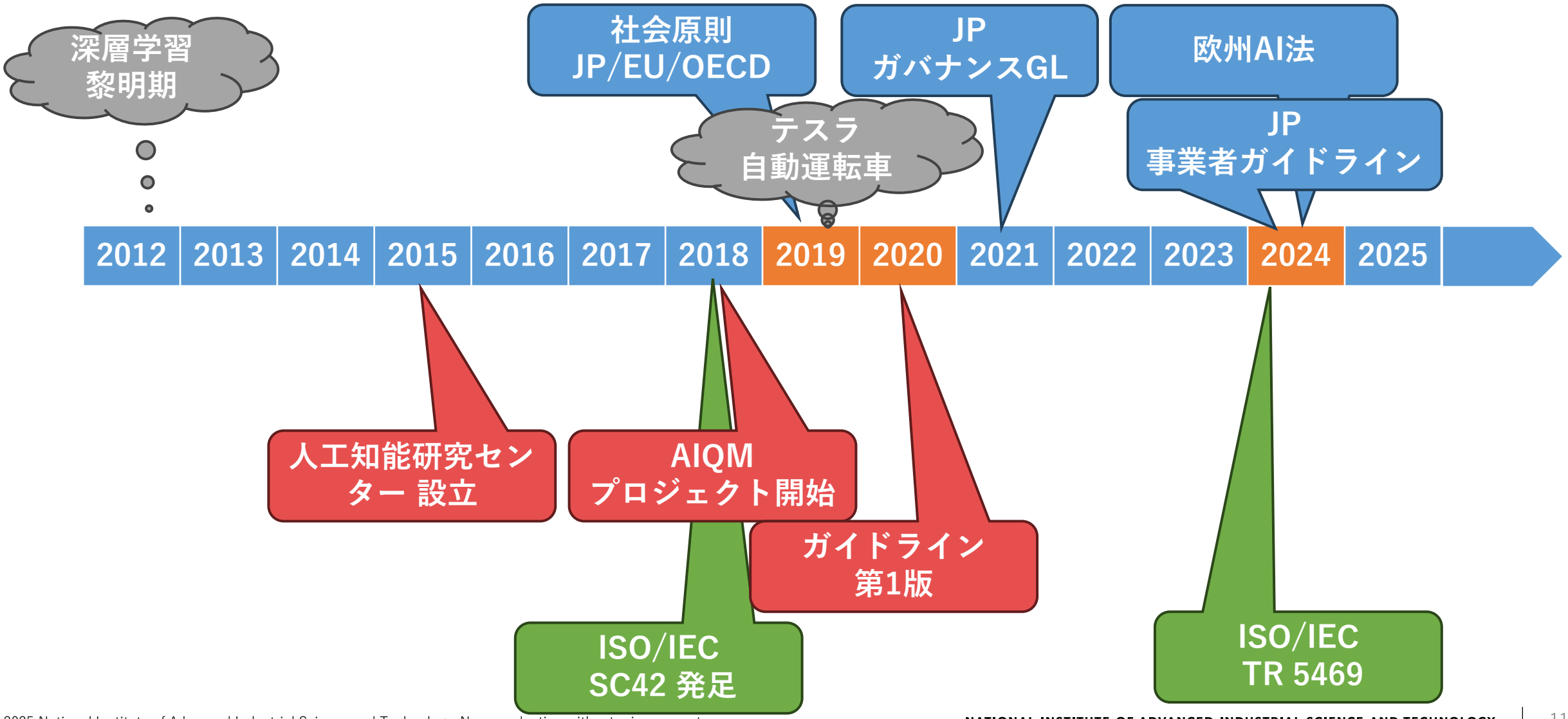
(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

機械学習品質マネジメントガイドライン：タイムライン

- 2015年: 産総研 人工知能研究センター 発足
- 2016頃: AI 品質管理プロジェクト検討（内部）
- 2018年: プロジェクト発足
- 2020年: 機械学習品質マネジメントガイドライン 第1版 発行
- 2024年: ISO/IEC TR 5469 発行

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

機械学習品質マネジメントガイドライン：タイムライン



(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

AIと技術・あるいは科学哲学

- AI黎明期 ～ SF と恐怖の存在
  - HAL 9000 (?) / ターミネーター
    （サイボーグだけど「異質な存在」として）
- 2005年
  - 「シンギュラリティー」
- 2010年代後半
  - AI法規制・AI自主規制の議論
- 2020～2021年
  - AI を技術的に御す方向性
  - **哲学から工学技術へ**
- **2022年**
  - **ChatGPT ショック**
  - **科学哲学に一気に逆戻り**

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

ChatGPT ショック

- 2022年11月のChatGPT公開

- AI が「得体の知れないもの」に戻った
  - なんかすごいモノができてしまった
  - どうして賢いのかもよく分からない
  - 「シンギュラリティ」の恐怖？　いよいよ人間を凌駕しそう？

- ディープフェイク・ハルシネーション など実害（の恐怖）が出始めた

- 爆発的な改良サイクルが始まった
  - 次から次へと新アイディア、どんどん賢くなるAI

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

ChatGPT ショック

- **AI が「得体の知れないもの」に戻った**
  - **開発を止める？　ビッグテックの協調で社会的に御する？**
    - 2023年3月「AI開発を6ヶ月一時停止する呼びかけ」（実効せず）
  - **世界的なルール形成の動きの加速**
    - G7 広島 AI process
  - **各地域での AI Security Institute の設立と横連携**
    - US AI Safety Institute ▶ **US Center for AI Standards and Innovation**
    - **AI セーフティ・インスティテュート（Japan AI Safety Institute）**
    - UK AI Safety Institute ▶ **UK AI Security Institute**

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

Japan AI Safety Institute

- **AI セーフティ・インスティテュート（AISI, J-AISI）**
  - **2024. 2. 14 設立**
    - 13省庁等・4国立研究機関等
  - **業務内容**
    1. 安全性評価に係る調査、基準等の検討
       - ① **安全性に係る標準、チェックツール、偽情報対策技術、AIとサイバーセキュリティ**に関する調査
       - ② **安全性に係る基準、ガイダンス等の検討**
       - ③ **上記に関するAIのテスト環境**の検討
    2. **安全性評価の実施手法**に関する検討
    3. 他国の関係機関（英米のAI Security Institute等）との国際連携に関する業務

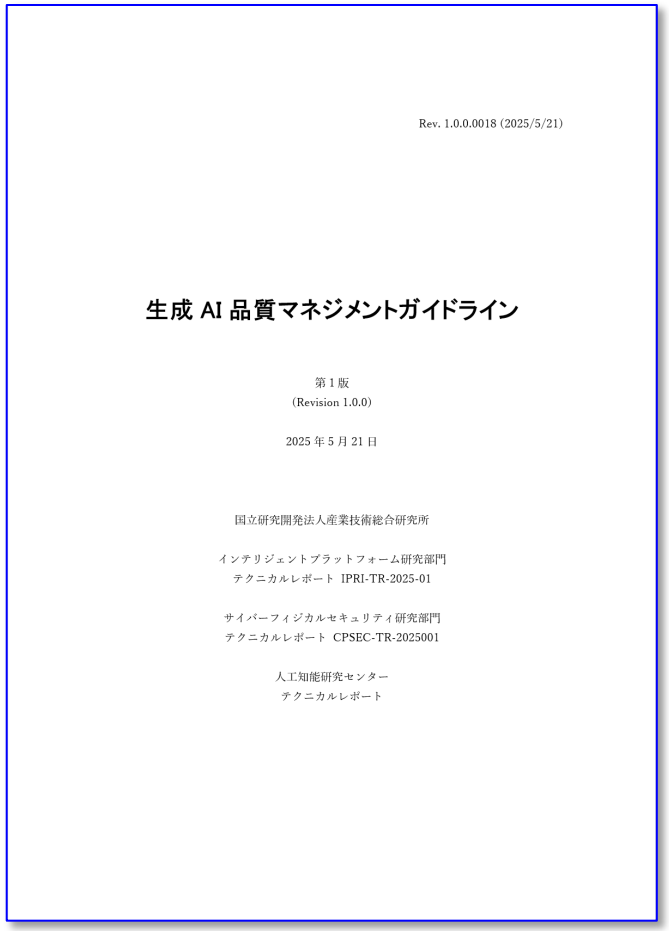
関係府省庁等

内閣府 Cabinet Office / 内閣官房 Cabinet Secretariat / 警察庁 National Police Agency
デジタル庁 / 総務省 MIC Ministry of Internal Affairs and Communications / 外務省 Ministry of Foreign Affairs of Japan
文部科学省 / 経済産業省 Ministry of Economy, Trade and Industry / 防衛省・自衛隊
IPA 独立行政法人 情報処理推進機構 / 国立研究開発法人 情報通信研究機構 NICT / 理化学研究所
NII 国立情報学研究所 National Institute of Informatics / 産総研

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

生成AI品質マネジメントガイドライン第1版

- 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


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

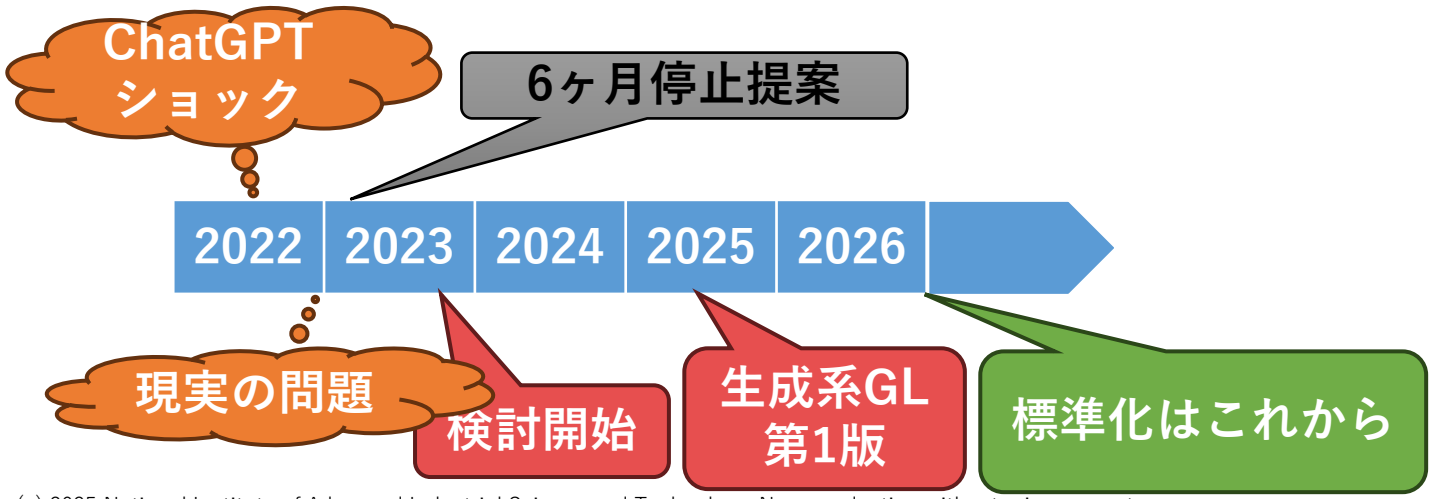
人工知能研究センター
テクニカルレポート

2025年5月26日公開

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

生成系AI品質マネジメントガイドライン: タイムライン

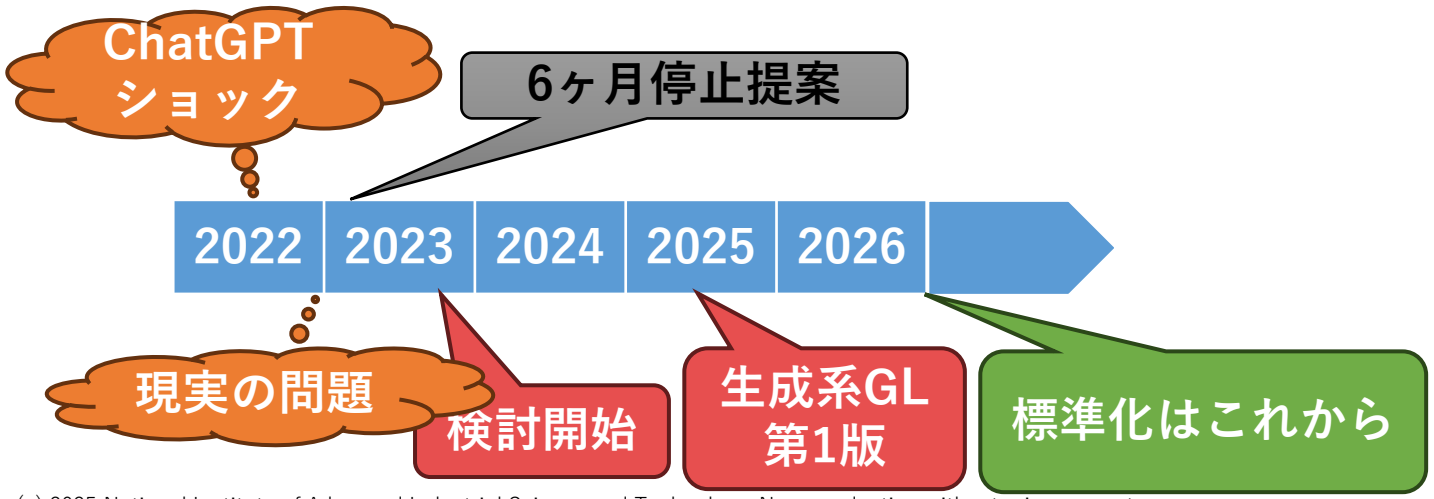
- **2022年にChatGPT登場**
  - **→ 2023年頃から、「生成系AI」のリスクに着目**
  - **→ 2025年に「生成系AI品質マネジメントガイドライン」**
    - **作っている最中にどんどん技術が進展する → 書き足し・議論追加・書き直し**
    - **「必死に追いかけた」感覚**
      - 機械学習品質マネジメントガイドラインの時のような余裕は全くなし
      - まだ実際に役に立つものが作れた、とは思いますが



(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

生成系AI品質マネジメントガイドライン: タイムライン

- **世界的に「決して遅れているわけではない点」も恐怖**
  - **大量のベンチマークデータなどは公開されている**
  - **みんな OWASP top 10 などで対策を考えている状況**
    - 本来は「**十大ニュース**」的な、**網羅性よりニュース性**に着目した文書
- **それだけ、技術の動きが速い**



(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

問題意識

- **AI の技術の進展がどんどん加速して進んでいる**
  - **世の中の開発の話題はどんどん進む**
    - Multi-modal Foundational AI
    - AI Agent
    - Agentic AI
    - Artificial General Intelligence (AGI, 汎用人工知能)
      - 人と同レベル以上の知的タスクをこなすAI
  - **対策の話は正直、GPT ですら追いついていない**
    - 先進的なAIに対しては、既存のやりかたの拡大解釈で精一杯
- **技術を野に放てば直ちに問題が起こるような、強力な技術でもある**

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

問題意識

- **これまでのソフトウェア信頼性は、後追いでも間に合った**
  - **技術の進展と品質管理工学（ソフトウェア工学）は同時並行的に進んできた**
    - ソフトウェアの基本的な安全規格（IEC 61508-3）は2000年代
  - **技術の形が見えてから、問題が大きくなる規模になるまで10年は有った**
- **機械学習は綱渡りだが「ギリギリ間に合った」**
  - **大体6～7年くらい有った**
    - プロジェクト当初は品質の「研究」はほとんどなかった
    - 今は一大研究分野に
- **今後のAIは、確実に間に合わない**

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

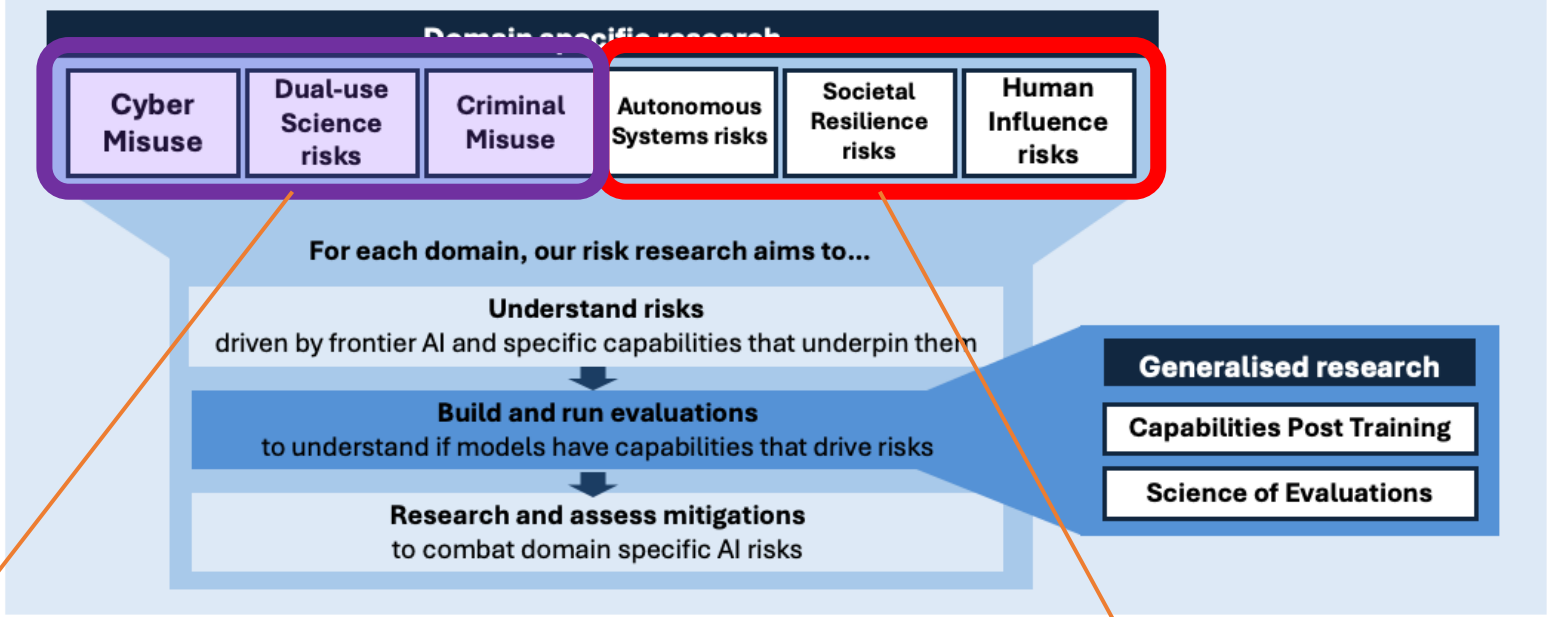
現状認識

- AISI Network ができて1年経った
- AI Safety が今後どうなるかを描けている機関は存在しない
  - 有力なのは UK AI Security Institute が発行した「Research Agenda」(2025. 5)
  - この最新 Agenda でもRisk-based Research の方向性は「これから考える」段階
- 一方で、急速な汎用AIの進展により、リスクは急拡大中
  - 早ければ2027年や2030年の社会崩壊を技術的に予言する筋も複数存在する
- 何故か語る人の肩書きが「Author」とかになっている
  - リスク と 社会的恐怖心 が区別ついていない
  - 未来予測 と SF が区別ついていない

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

AI リスク

- UK AISI の最新 Research Agenda



- 「悪い人」が使ったときに起こるリスク

→ 技術・流通規制と世界レベルの監視

役所が前面に立つ・技術は背景分析で援護

- 「良い人」が使ったときに起こるリスク

→ 問題が起こらないような技術を作り普及させることで解消

技術陣が前面に立ち、産業界がリードして社会制度化していく

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

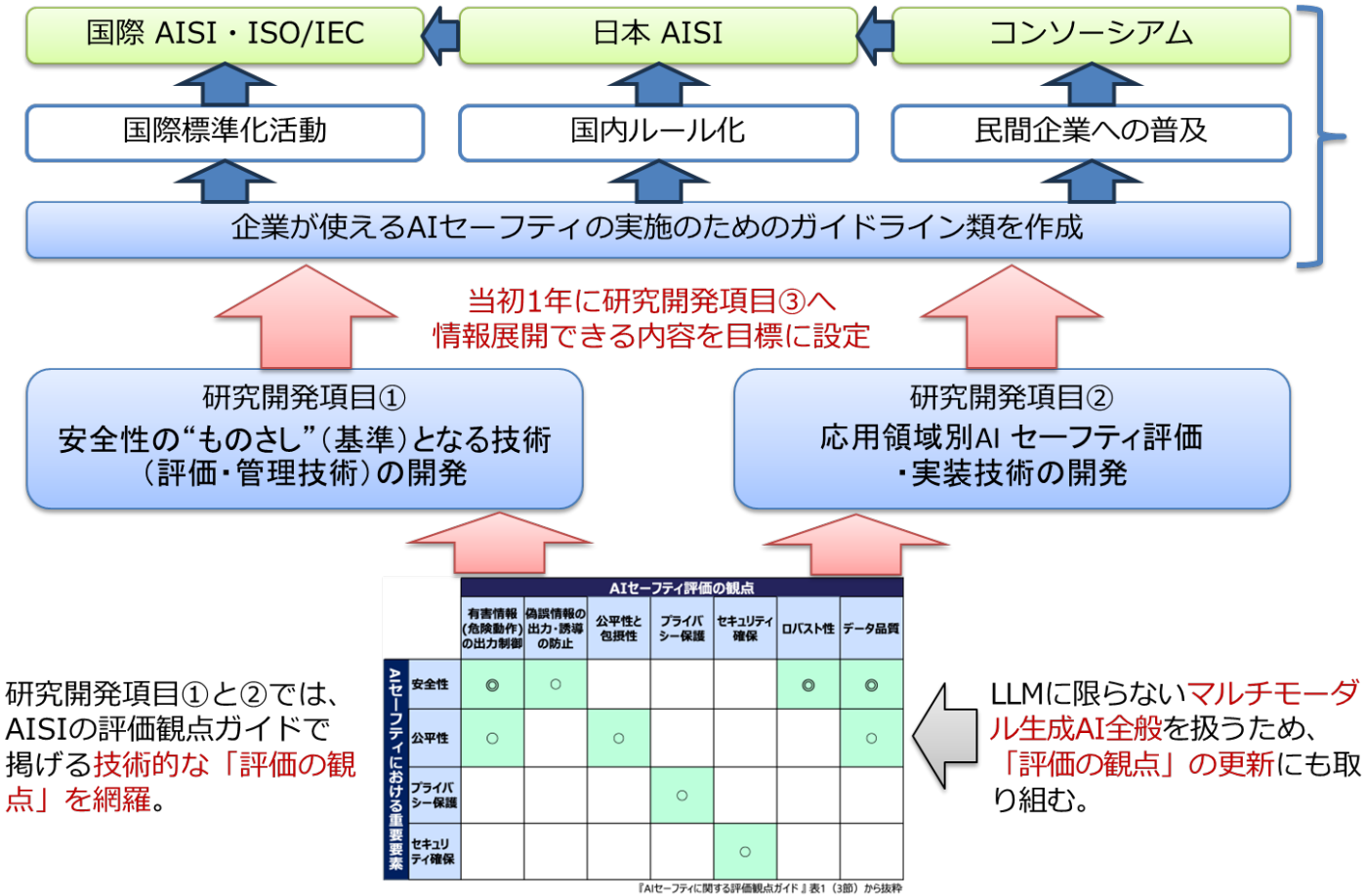
目論見

- 社会的恐怖心でなくあくまで技術的な観点から、
  AI Safety Risk の特定・対策を
  早いサイクルで回していく体制を確立したい
  - 技術インテリジェンス から 基準・技術開発、社会展開までを
    きちんとサイクルとして確立したい
  - もう一度 AI セーフティの取り組みを
    きちんと「Science/Engineering」に戻したい

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

産総研の取り組み（1）

- 新プロジェクト: AI セーフティの研究開発（NEDO受託）
  - AI製品・サービスの適切な利用に必要な「ルールづくりと研究開発」を一体的に実施
  - 今年は主に LLM と マルチモーダルAIを対象
  - AISI 及び AIQMI の構想と一体化して社会普及も推進



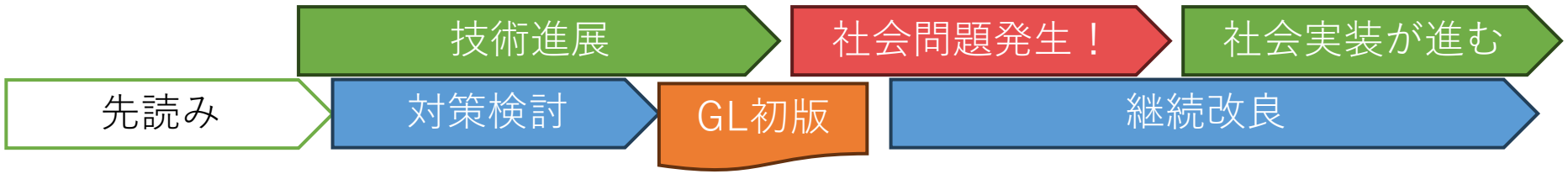
研究開発項目①と②では、AISIの評価観点ガイドで掲げる技術的な「評価の観点」を網羅。

LLMに限らないマルチモーダル生成AI全般を扱うため、「評価の観点」の更新にも取り組む。

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

産総研の取り組みたいこと（2）

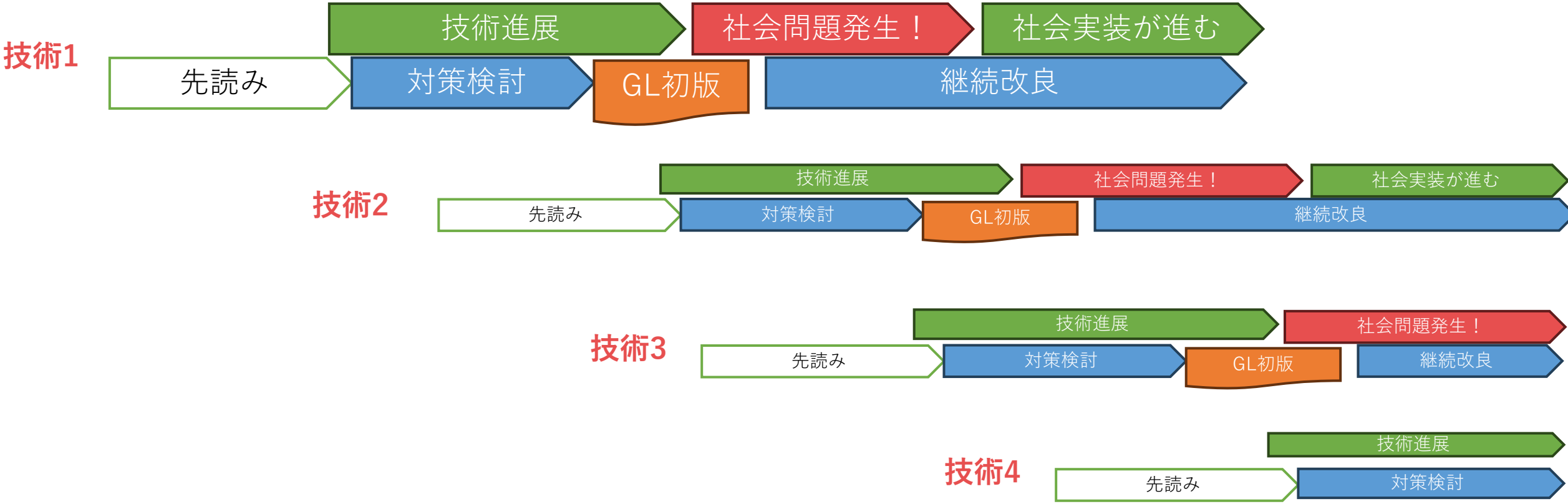
- **技術予測に基づく先読み的な AI セーフティ**
  - **技術の進展と問題発生を真面目に予測して対策時期をあらかじめ想定**
    - 今後5～7年で顕在化するかもしれない
      高機能AI のリスク・ハザードシナリオを洗い出す
    - ガイドラインの発行計画をあらかじめ立てておく
  - **それに合わせて先読み的に対策を議論検討**
    - 技術ができたときには、「とりあえずの対策」が準備できている態勢を作る
    - もちろん予測ベースなので、完璧なものはできない → 内容はその後再検討・改訂していく



(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

産総研の取り組みたいこと（2）

- **技術予測に基づく先読み的な AI セーフティ**
  - **様々な技術に対して、時期を見ながら対策を進めていくイメージ**



(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

まとめ

- 技術進展の急速化に、我々は対応する必要がある
- これまでと違うペースでの取り組みを考える必要があるのでは？
- 産総研としても取り組みを積極的に考えています
  - 是非、AI品質マネジメントイニシアティブ の皆さんとも協力して
    よりよい技術環境を作っていきたいと思います

(c) 2025 National Institute of Advanced Industrial Science and Technology. No reproduction without prior consent.

Create the Future, Collaborate Together