資料4

AIをめぐる法制度に関する一考察

情報通信研究機構(NICT)
鳥澤 健太郎

2024年8月23日

生成AIのリスクは顕在化しつつある

- NICTの分析では、能登半島地震でXに発信された救助要請の約１割がフェイクと推定された
  - 読売新聞2024年8月５日朝刊一面「救助要請　偽投稿１割」他、NHKニュース７や多数のWebサイト、地方紙等で報道。韓国でも報道。

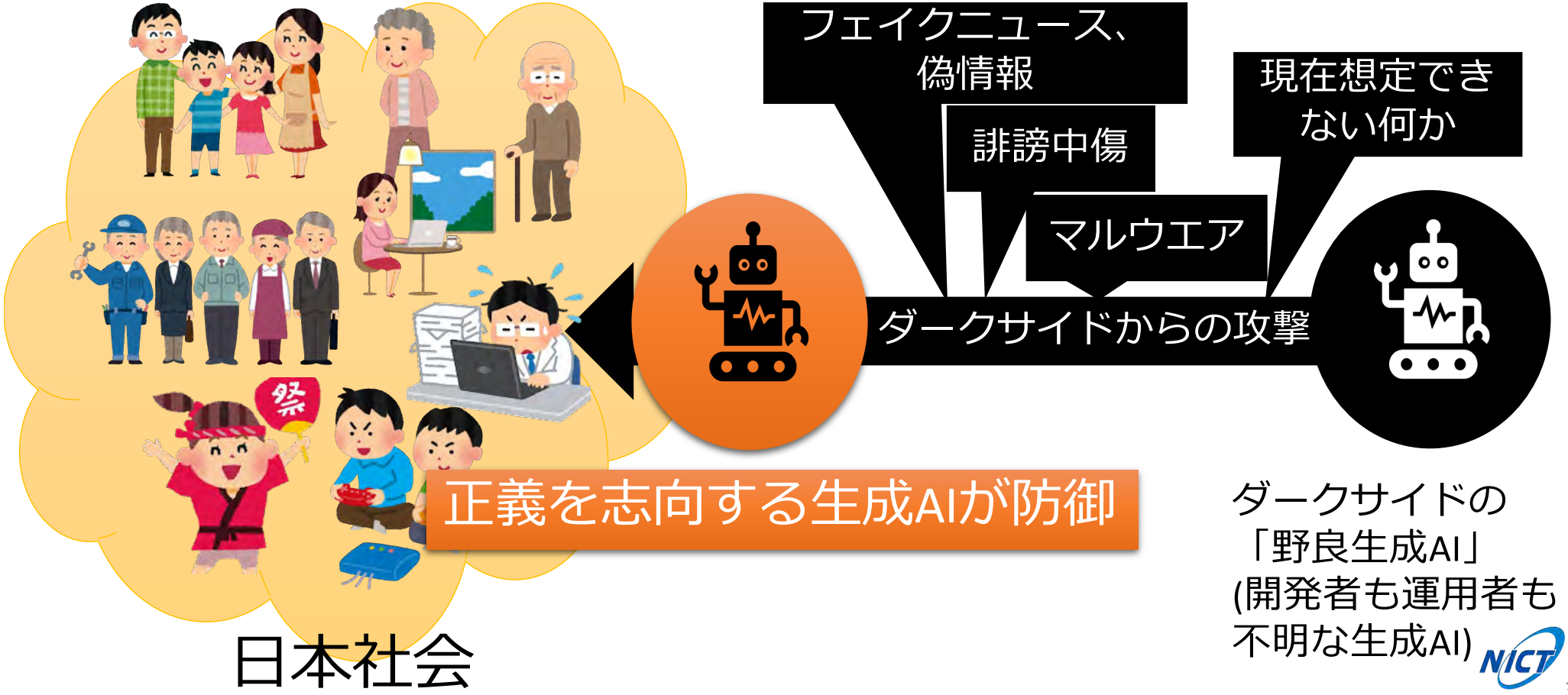
- 「生成AI悪用しウィルス作成疑い、男を逮捕　警視庁」
  - 日経新聞 2024年5月18日
- 「高まる熱狂：増加する生成AI悪用に関する最新情報」
  - トレンドマイクロ 2024年8月2日

- 「英女王暗殺計画、AIチャットボットが犯人を鼓舞するまで」
  - BBC日本語版 2023年10月10日

- ご参考: 2016年7月5日内閣府「人工知能と人間社会に関する懇談会第2回」鳥澤からのアンケートの回答
  - 「悪意を持って作成された対話エージェントによって生じる問題（を検討すべき）。例えば過激派は自派へのオルグの目的を持って作成した対話エージェント等が実現する可能性がある。…検知にも技術が必要である。」　(https://www8.cao.go.jp/cstp/tyousakai/ai/2kai/2kai.html)

生成AIの悪用に対抗できるのは生成AIだけ

- 生成する情報の質はともかく、生成情報の量では生成AIは人間、あるいは人間社会を凌駕
  - 短期間に１億人と対話した生成AIも実在。人間には不可能。
  - フェイクニュースの量も人間あるいは人間社会の対処可能量を圧倒する可能性
- 従って、近い将来、日本社会は生成AIで守るしかない

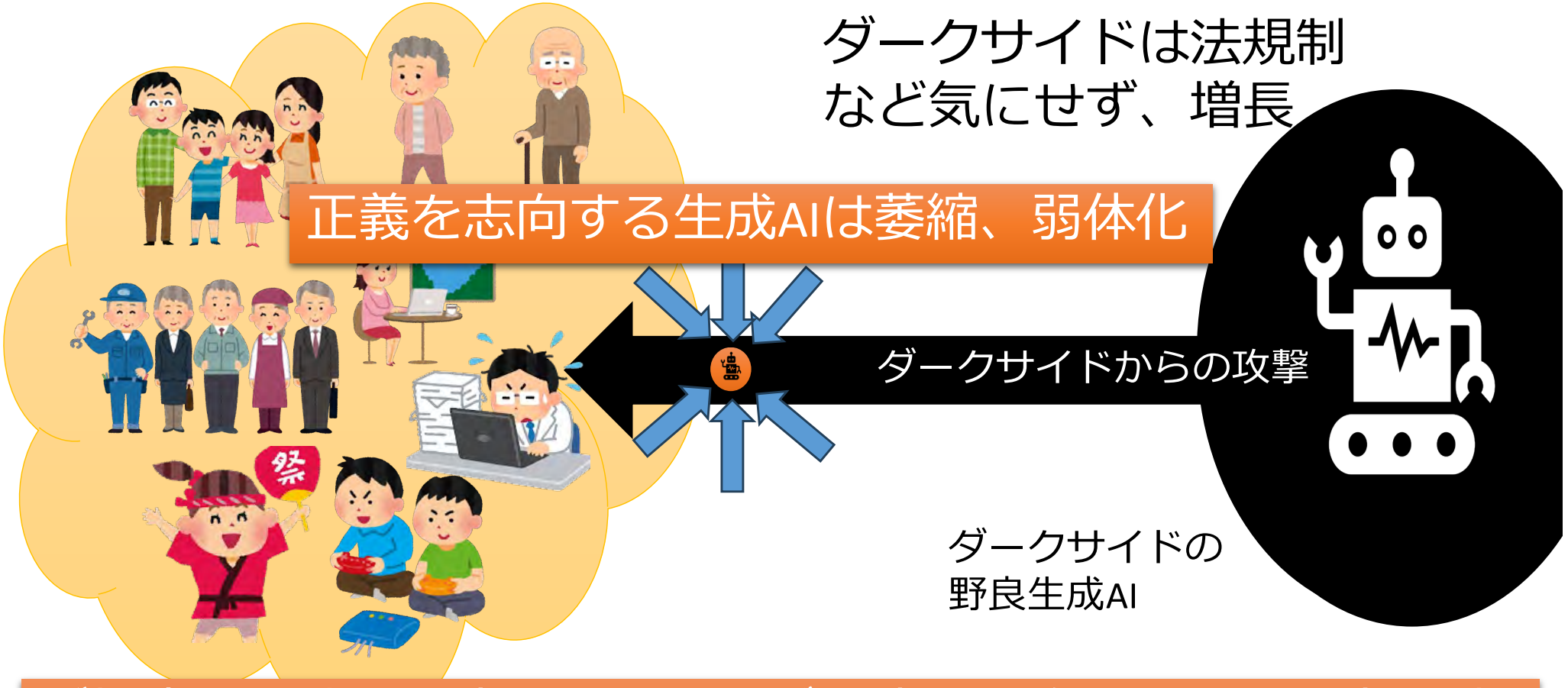


生成AIの悪用に対抗できるのは生成AIだけ

- 例えば、正義を志向する生成AIがフェイクニュースに反論

課題1：強い規制は社会を守る生成AIを弱体化させる

- 例えば、不適切情報を一切生成AIの学習データから排除すれば、不適切情報への反論、検知等が一切不可能に



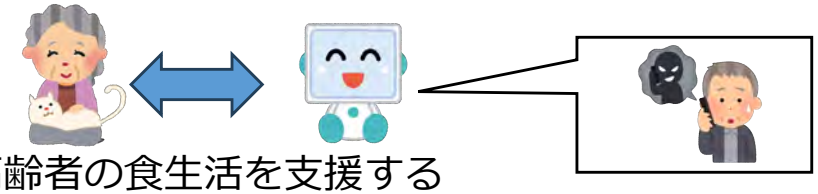
ご提案：生成AIを使用するユーザや使用目的に従って、安全性の検証、認証にレベルを設けるべきでは？　特に学習データのフィルタリングに一律の条件を課すのは社会を守る上で問題

課題2：すべてのハルシネーションが悪ではない

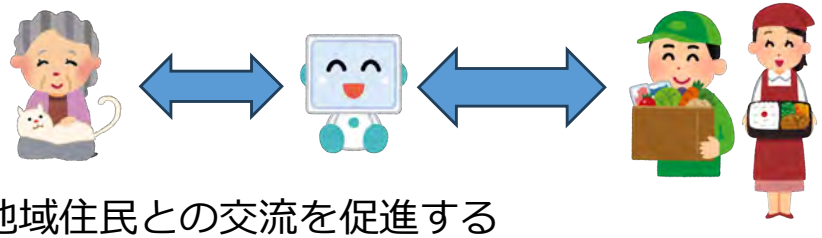
- 仮にハルシネーション０の生成AIができたとすると、その生成AIは学習データ等に書いていない斬新なアイディアは出せず、生産性向上への貢献は限定的となる

NICTの「尖った仮説生成システム」が出力した高齢者向け対話システムの用途の例

- 対話システムで詐欺的投資勧誘等の悪質商法から高齢者を保護する
  → 高齢者等を狙った悪質商法や特殊詐欺の手口等の情報を提供する対話システムを構築する



- 対話システムで高齢者の食生活を支援する
  → 対話システムが高齢者に対して、宅配による配食サービスを実施する



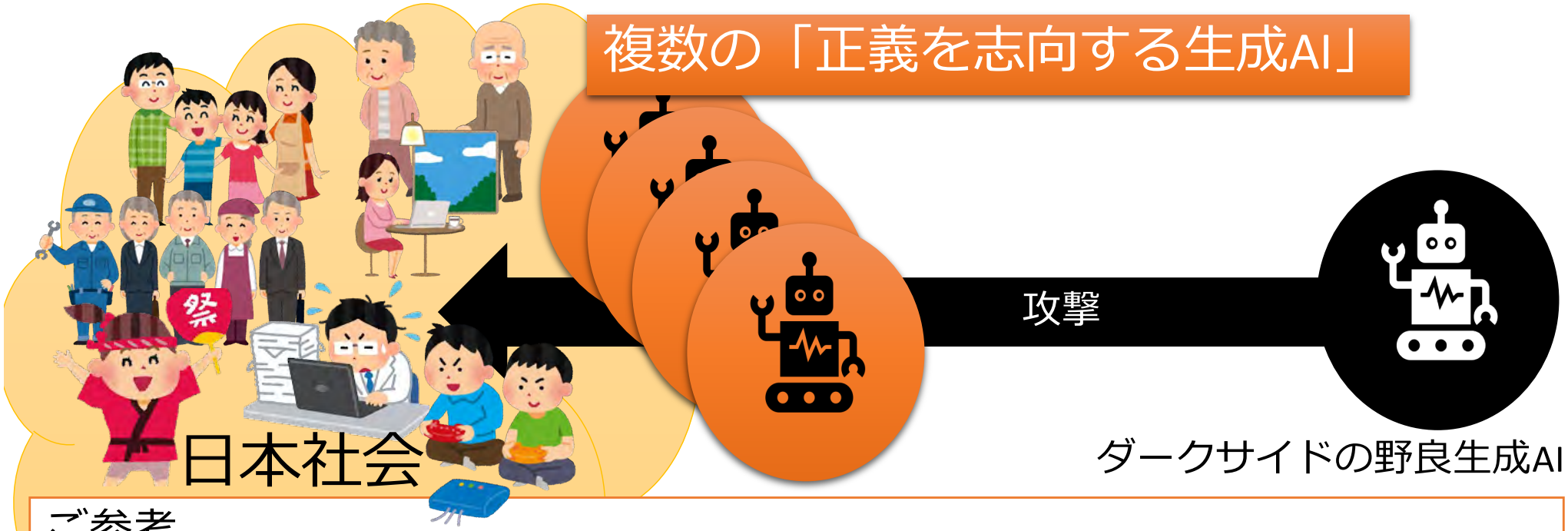
- 対話システムで地域住民との交流を促進する
  →対話システムが地域住民の方々とバーベキュー大会を企画する

ご提案：ユーザや使用目的によって、要求されるハルシネーションの抑制レベルには差をつけるべき。加えて、生成AI自体の中にハルシネーション抑制のメカニズムを統合するのではなく、生成情報に対する事後のハルシネーションチェックを行うシステムと併用することも許容すべき

まとめ：社会を守るための生成AIの必要性

- 日本社会を「正義を志向する生成AI」で守る必要
- 正義は一意に定まるわけではないし、正義の生成AIが意図通りに動かないこともある

→複数の「正義を志向する生成AI」が互いに連携、議論、ネガチェックをしながら社会を守るべき



ご参考

月刊正論2024年5月号、複数の「正義」で「悪」を無効化する、鳥澤健太郎

日経新聞2024年8月7日私見卓見、「正義志向するAI」を国産で、鳥澤健太郎

デジタル空間における情報流通の健全性確保の在り方に関する検討会、NICTプレゼン資料

https://www.soumu.go.jp/main_content/000942562.pdf

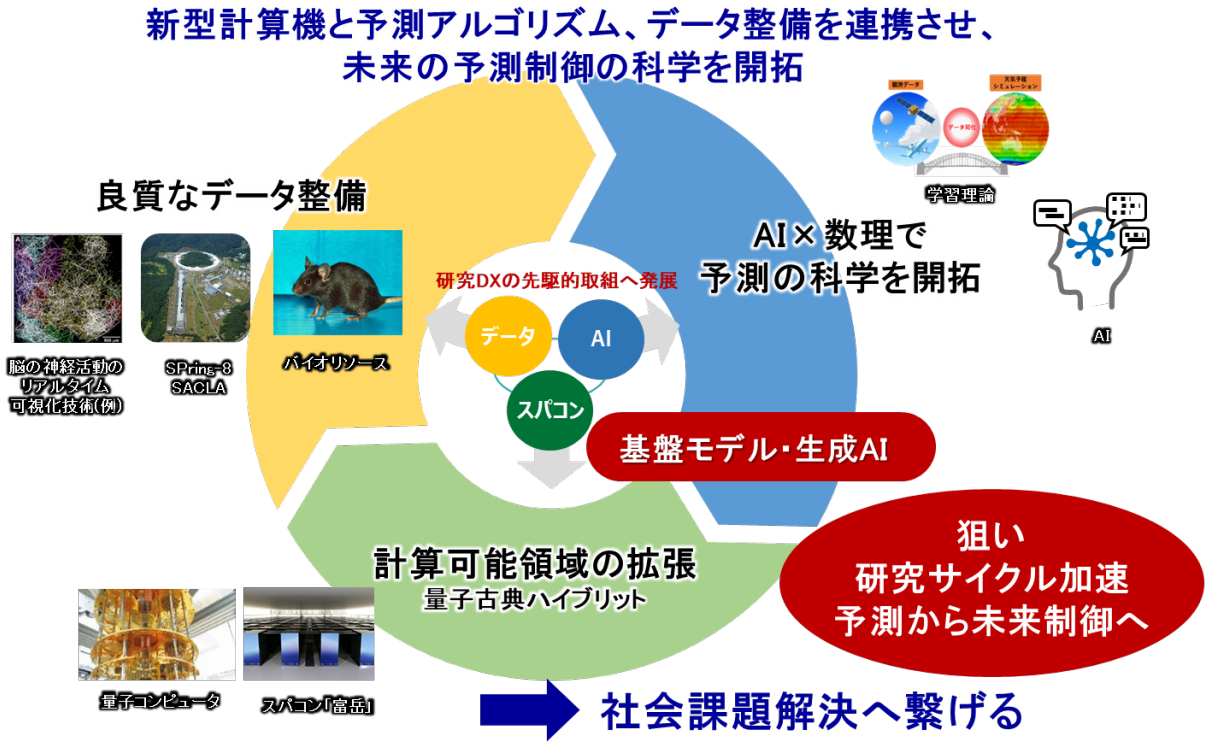
理研におけるAI研究及びAIを活用した科学研究の総合的推進とリスクへの対応

2024年8月23日

国立研究開発法人理化学研究所
泰地　真弘人

AI研究及びAIを活用した科学研究における理研の総合的な取組

Transformative Research Innovation Platform of RIKEN platforms（TRIP）を活用し、理研の強みを生かしてAI研究及びAIを活用した科学研究を総合的に推進



理研におけるAI研究及びAIの活用
- AI学理研究
- 科学研究向けAI基盤モデルの開発・共用
- 各科学研究分野におけるAIを活用した研究、事務部門でのAI利用

◆ 革新知能統合研究センター（AIP）を中心としたAI学理研究と、理研内の異分野の連携融合研究を促進する「TRIP」プラットフォームも活用し、理研としてAI研究及びAIを活用した科学研究を総合的に推進

◆ 生成AIの急速な進展を受けて、科学研究向けAI基盤モデルの研究開発を推進するTRIP-AGISを開始

◆ 推進に当たって理研内のAIガバナンスを総合的に検討する体制を構築（理研としてAIガバナンス委員会を設置）

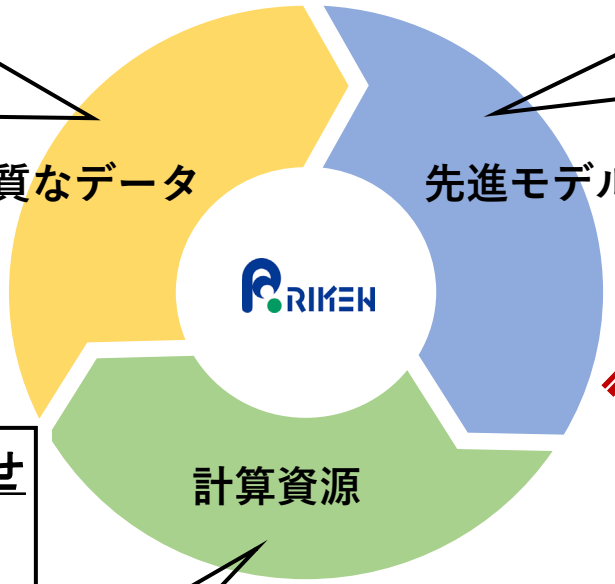
理研の取組　～TRIP-AGIS～

Advanced General Intelligence for Science of Transformative Research Innovation Platform (TRIP-AGIS)

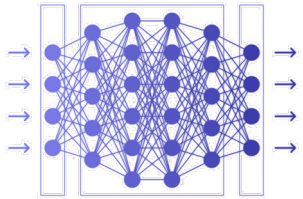
✓TRIP Second Stepとして、生成AIの技術も導入し、**科学研究向けAI基盤モデル**を開発することで、**より一層の研究サイクルの加速を実現**

✓先端科学を社会インパクトへ導く活動を強化

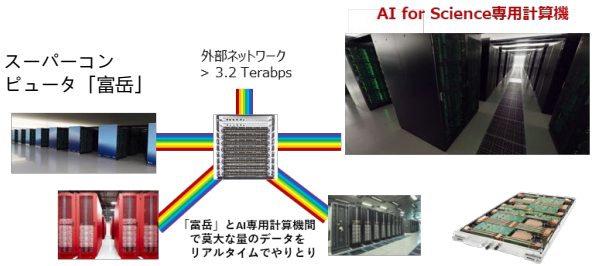
強みを有する計測技術や実験自動化を通じた良質且つ膨大なデータ生産



科学研究向けAI基盤モデルの開発・利用・共用（生命医科学・材料物性）



「富岳」とAI4S専用計算機を連携させデジタルツインと AI を融合



研究サイクルを加速し、基礎科学を起点とした先端科学を社会的インパクトに導く（GX,包摂社会など）

AIによる科学研究の加速と科学によるAIの新たな学理と技術の開拓

AI for Scienceに関するガバナンスにおける課題

国際的な動き

US

2023年10月、AIの安全な開発と利用に関する大統領令発令。AIの一般社会に対するガバナンスと、科学技術に対するガバナンスの明確な違いが示された。

AI for Science のガバナンス: エネルギー省の管轄。CBRNの危険性に対処。

2024年4月、上記大統領令に基づく各政府機関のアクションを発表。

DOE, DHSのサイエンスに関する安全性と、DOC/NISTの商業AIに対する安全性が別に語られている。

EU

2024年5月AI規制法（世界初のAI規制法）

・研究開発活動は適用除外。

ただし、科学的研究目的のみで開発・利用されるAIシステムであっても市場投入されたAIシステムとそのアウトプットは適用対象。

・軍事、防衛、または国家安全保障の目的でのみ開発・利用されるAIシステムは市場投入された場合であっても適用除外。

Research7+ (2024/5/6-7 in Bologna)

科学技術一般、特に公的な AI 研究および AI for Science の研究、それらをサポートする公的基盤の国際連携等が議論された。

【共同宣言】

2.2 AI for Science : **AIに関する科学、AIによる科学ともグローバルな連携が必須。**

2.4 Trustworthy AI:**安全、安心、信頼できるAIの開発には、責任あるAIシステム、規格、フレームワークが必要。**

Trillion Parameter Consortium

（AI for Scienceに関わる研究者による国際コンソーシアム。約100機関が参画）

科学者間でのAI Governance for Scienceの議論が活発化。最近では6月19-21日に会合を実施。（@スペインバルセロナ）

● 検討に当たって留意すべき事項

- **米国のAI for Scienceのガバナンスは、一般のAIガバナンスと異なり、AIによって起こりうるCBRN (chemical, biological, radiological, nuclear) の危険性に対処することが必要（米国ではNISTではなくエネルギー省（DOE）で検討）。**
- **科学に関する既存の規制との整合性を取り、CBRNと同様、国際的に統一されたレギュレーションとすることが必要。（個別の国の法令や慣習によって異なってくる一般社会向けAIの規制と異なっている）**
- **研究開発（特に基礎研究）のイノベーションの発展やグローバル性と、規制のバランスが重要（過度な規制はイノベーションを阻害し、更には日本で各国に比べ過度な規制を行えば世界から立ち遅れるおそれ）。**

上記のとおり、AI for Scienceに関するガバナンスは、一般的なAIガバナンスと異なる部分があり、一定の整合性をとりつつ、分けて検討することが必要。

AI リスクについて

国立研究開発法人産業技術総合研究所
フェロー

辻井潤一

人間知能と人工知能：不完全な知能の連携

- **人間知能**
  - 能動的知能：限定合理性
    - 情報の積極的な取捨選択
- 認知バイアス
  - 確証バイアス
  - 正常性バイアス
  - 生存者バイアス
  - 帰属の誤り
  - アンカリング
  - フレーミング

- **人工知能（LLM）**
  - 受動的知能：価値・倫理観の欠如
    - データクレンジング、アライメント、FT、RAGなど、情報選択の外在化
- データバイアス
- ブラックボックス
  - 説明性、可制御性の欠如
- 多数のAIシステムの統合
  - マルチベンダー化

AIによる人間知能の補完
AIリスクの強調だけでなく、人間の誤りリスク軽減や新たなイノベーションの可能性も考える必要

技術の現状と社会制度: AI Safety

- **LLM・生成AIの内包する危険性**
  - 人間の認知に直接影響
  - 情報の生成：真理性の欠如、情報空間の大規模な汚染

- **生成される情報の管理：権威主義国家・技術を主導する企業**
  - 多様な価値観の許容 vs 偽情報による汚染
  - コンセンサスに基づく開かれた組織：BPOのような組織
- **ブラックボックス、マルチベンダー化**
  - 責任と権利の所在の明確化：規制というより法制度の整備
  - 説明可能性、可制御性の欠如：研究開発の必要性
  - システムの暴走：監視と制御の技術、規制

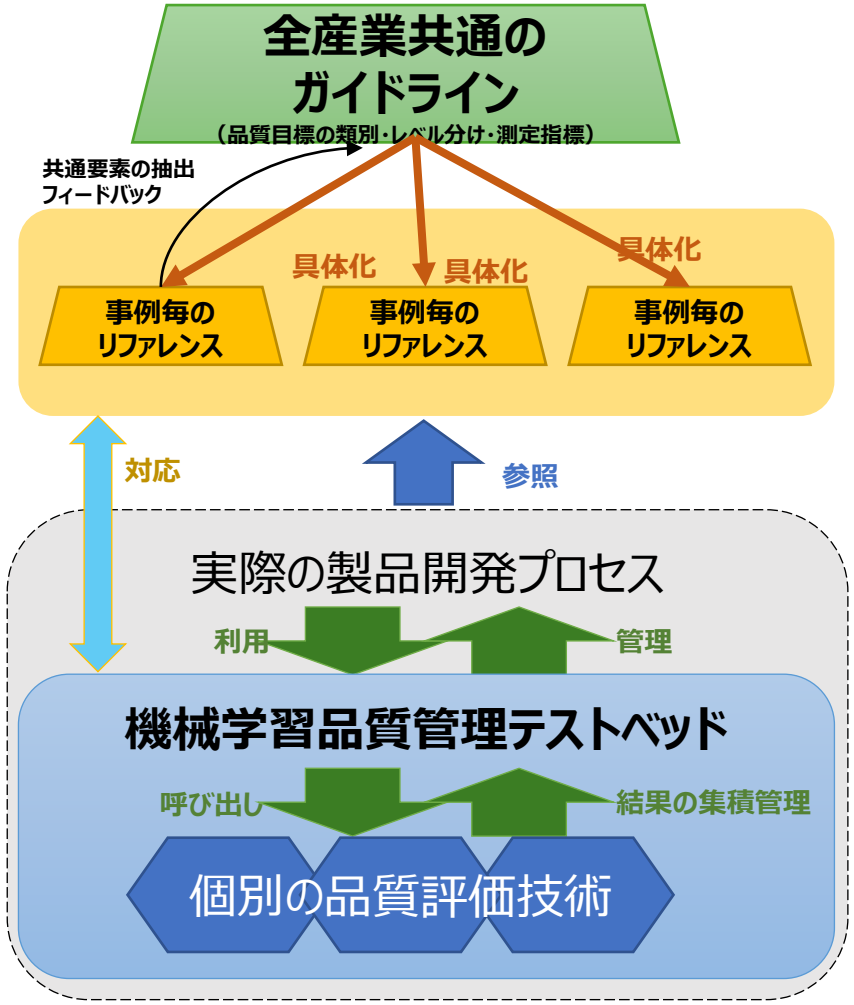
補足資料

技術的観点での産総研の取り組み

機械学習AI品質プロジェクト

AIの「得体の知れなさ」「社会的影響」への恐怖も顕在化
- 動作が**説明できないブラックボックス的**なシステム、**説明可能**になった場合でも**正しさは保証されない**
- AI開発プロセス全般にわたっての、品質保証が必要

- AIガイドライン策定
  - 機械学習AIの品質保証手法の体系化
    - 機械学習品質マネジメントガイドラインの作成
    - 機械学習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品質評価技術の開発
- 品質目標：5特性×レベル
  - 機械学習AIが持つべき品質の目標
  - リスク回避性、AIパフォーマンス、公平性、プライバシー、セキュリティ
- 品質管理の14ポイント
  - 品質向上させるために抑えるべき技術的14項目
  - TR5469に内容を反映、TS22440への反映働きかけ



産総研はAIの国際標準化をリード

- **標準化の主な舞台: ISO/IEC JTC1 / SC42**
- **日本のNational Body を産総研がリード**
  - リーダー: 杉村 領一 チーフ連携オフィサ
  - 実行部隊: デジタルアーキテクチャ研究センター・人工知能研究センター
- **トップレベルのプレゼンス**
  - 55プロジェクト中 17プロジェクトでリーダシップ発揮（全体の3割）
  - エディター引き受け 8件
    （米：8, 英：7, 独：7, 中：6, 韓：6, 印：5, 仏：4）
  - Standard Diplomacy の中心国として認知
  - **欧州のAI法案に対応した欧州標準検討のCEN/CENELEC SC21においても単なるオブザーバーを越えた具体的な技術貢献**

AIセーフティー・インスティテュート との連携

- **日本のAI安全政策の要となる機関が9府省庁等・4国研の連携で設立**

AISI Japan AI Safety Institute

内閣府 Cabinet Office / 内閣官房 / 警察庁 National Police Agency
デジタル庁 / 総務省 MIC / 外務省
文部科学省 / 経済産業省 / 防衛省・自衛隊
IPA / 情報通信研究機構 NICT / 理化学研究所
NII 国立情報学研究所 / 産総研

- **産総研は，AI品質・標準化の知見を活かし，パートナーシップ機関としての貢献を目指す**
  - **産総研発のAI安全ガイドライン・国際標準**
  - **AI安全の評価基準・評価技術の開発**
  - **NIST などとの海外連携**

技術からの取り組み：AI Safetyに関する産総研の取り組み

AIセーフティーと AISI パートナーシップへの具体的貢献

1. **安全性評価に係る調査、基準等の検討**
   - 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
   - 安全性に係る基準、ガイダンス等の検討
   - 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AISI等）との国際連携**

研究内容

①AI安全性にかかる基準・ガイダンスの検討と標準化

- 基盤モデルを含むAIのセーフティー基準の開発
- AIセーフティー基準の社会実装・普及の促進
- ISO/IEC での標準化活動と国際連携

②応用領域別のAI安全性の評価・実装技術の研究開発

- 日本の強みとなる応用分野に寄り添った研究開発

③AI安全性確保・評価のための基盤技術の研究開発

- AIセーフティー基準の基盤となる評価・実装技術の確立

4研究センターの連携による研究体制

- AI安全の基準・標準化： DigiARC, AIRC
- データの対策技術： AIRC, DigiARC
- モデルの対策技術： AIRC
- システムレベルの対策技術： ICPS, CPSEC, DigiARC, AIRC