理化学研究所でのAIセーフティ・AIセキュリティに関係した活動のご紹介

黒川　原佳（くろかわ　もとよし）

国立研究開発法人理化学研究所
副理事（情報セキュリティ技術統括）
情報統合本部情報セキュリティ・システム部　部長

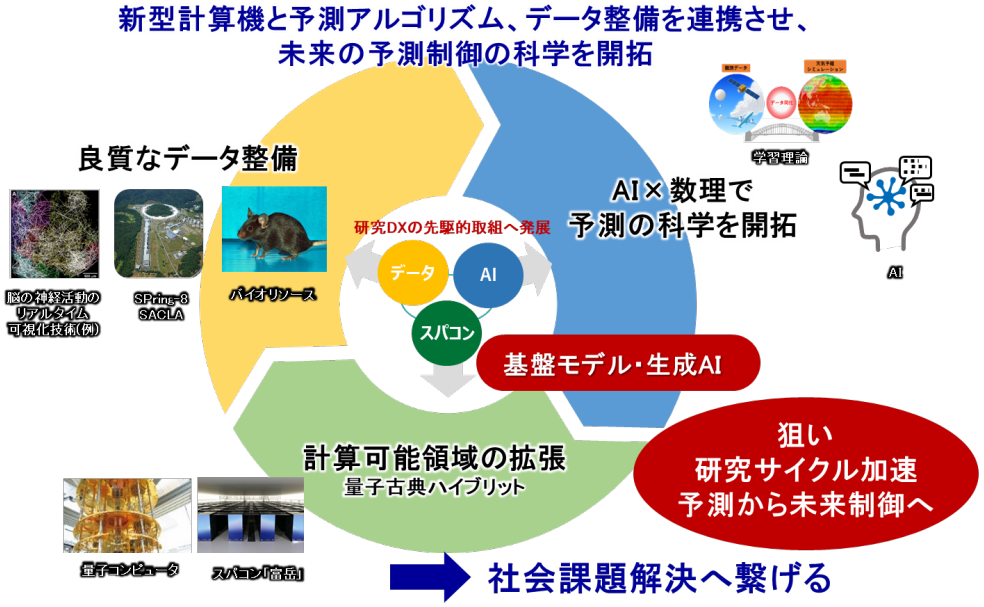
2026年5月21日

アジェンダ

- 組織でのAIセイフティ・ガバナンスに対する活動
- AI for Scienceの活動のご紹介
- 個別の活動のご紹介
- まとめ

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）

中長期目標・計画、年度計画におけるAIガバナンス

第5期中長期目標（2025年度～2031年度）

３．研究開発の成果の最大化その他の業務の質の向上に関する事項
　３．２　国際的な頭脳循環のハブ形成と研究環境に係る先進的な取組の実践
　　（３）社会状況・国際状況の変化への対応
　　（前略）
　　また、**科学研究における責任あるAIの研究・開発・推進を行うためのAIガバナンスなど、社会状況・国際状況の変化に対応するため、政府方針や社会の要請等を踏まえた体制整備を行うなどの適切な措置を行う。**

第5期中長期計画（2025年度～2031年度）

Ⅰ．研究開発の成果の最大化その他の業務の質の向上に関する目標を達成するためにとるべき措置
　２　国際的な頭脳循環のハブ形成と研究環境に係る先進的な取組の実践
　　（３）社会状況・国際状況の変化への対応
　　（前略）
　　理研における責任あるAIの研究・開発・推進を行うためのAIガバナンス等、新たな社会の要請等を踏まえた体制整備を行うなど、社会状況・国際状況の様々な変化に対して適切な措置を講じ、より強固な研究環境を確立する。

2026年度計画

Ⅰ．研究開発の成果の最大化その他の業務の質の向上に関する目標を達成するためにとるべき措置
　２　国際的な頭脳循環のハブ形成と研究環境に係る先進的な取組の実践
　　（３）社会状況・国際状況の変化への対応
　　（前略）
　　社会の要請等を踏まえた責任あるAIの研究・開発・推進を行うため、理研内に設置したAIガバナンス委員会にて、理研において必要となる科学技術発展へ貢献するAI（AI for Science）を中心に、理研のAIのガバナンスの検討等を行うとともに、政府方針や国際的な検討に即した対応を進めるなど、適切な措置を講じる。

最近の議論の概要

① AI関連の国際競争激化や日本でもAI法制・AI適正性確保指針の整備、AI for Science(AI4S)による科学研究加速プログラムや理研でもAI4S関連基盤整備や富岳Nextと様々な状況が進んでいる。

② 政府指針を参照しつつ、科学の発展に資するAIガバナンスを検討していくことが重要。

③ 人間中心の原則の堅持なのか?、、Agentic AI時代におけるAIとAI、AIと人間の関係を整理し、議論の範囲や実装を広げる必要がある。

④ 学際的・組織横断的にAIを連携させるには、共通ガードレールや統一的運用原則が不可欠であり、適切なガバナンス体制の整備が必要。

⑤ AI4Sの高度化にはデータは不可欠であり、法的制約のあるデータや個人情報の取り扱いに十分配慮しつつも、利活用するための検討が必要。

⑥ AI開発への過度な制約を課す動向や科学と直接的な関係のない議論が研究開発などに波及していることは懸念点である。

科学研究基盤モデル開発プログラム（AGIS）

Advanced General Intelligence for Science Program (AGIS)

✓TRIP Second Stepとして、生成AIの技術も導入し、**科学研究向け生成AI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

激変する社会は最先端科学による課題解決の加速を求めている

【深刻化する地球規模の課題例】

- GX加速へのニーズ：2050年までにカーボンニュートラル達成、循環経済への転換が国際的に浸透、技術大転換時代に突入

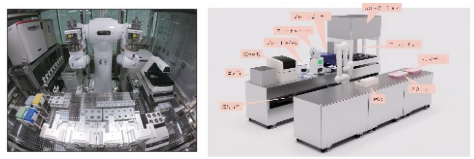
→脱石油化学のための資源循環型高分子化学

- 生成AI活用の急拡大：情報漏洩への懸念、分断・格差の助長等、社会への予期せぬ影響

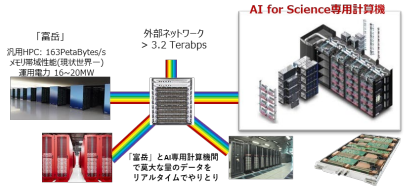
→AI for Science

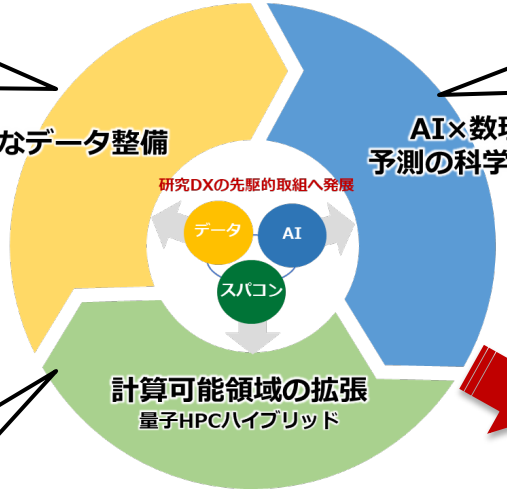
→社会に広く受け入れられる信頼性の高いAI（Science for AI）

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

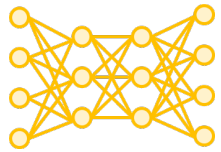


「富岳」とAI専用計算機を連携させ高度な学習・推論を実現





科学研究向け生成AIモデルの開発・利用・共用（生命医科学・材料物性）



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）

生成AIによる科学技術研究加速と科学による生成AIの新たな学理と技術の開拓

AGISの狙い

1. **Scalingへの対応**
   データ（量・質）/モデル/計算の規模拡大
2. **科学向け基盤モデルの開発と活用**
   マルチモダリティを重視
3. **実験/シミュレーション・解析の自動化**

→ **科学的成果の創出**
Grand challenge課題の解決

理研の取り組みは、現場の科学者を中心に、データ取得まで含めて行うAI for Scienceへの組織的な取り組みとして先駆的なもの

【参考】米国アルゴンヌ国立研究所でAI for Scienceの組織的枠組の検討を開始したのは2024年6月から
https://www.newswise.com/doescience/argonne-to-support-new-ai-for-science-projects-as-part-of-the-national-ai-research-resource-pilot

AGIS各プロジェクトの概要

✓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するため、令和6年度に開始した先駆的な取組

科学向けAI開発　ターゲット領域　→理研の強みを生かせる2領域を選定

泰地真弘人
(理研AGIS)

生命・医科学基盤モデル開発

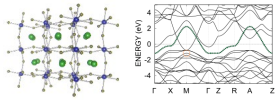
分子から個体・基礎から医療まで、多様な階層のモデル開発

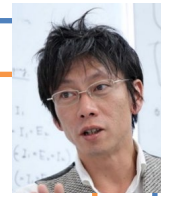
＜理研の強み＞
- ✓ 多様な階層における高い計測技術（例：細胞の刺激に対する動的応答のマルチモーダルデータを取得可能）

密接に連携

材料・物性基盤モデル開発

固体物性・高分子、加工技術等のモデル開発





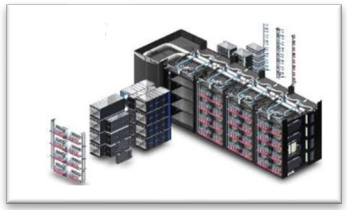
吉田亮
(理研AGIS/統数研)

＜理研の強み＞
- ✓ 世界最大級の高分子計算物性データベースを用いた計算機実験の自動化技術など

松岡聡
(理研R-CCS)

計算基盤

AI専用計算機による高度な学習・推論



＜理研の強み＞
- ✓ 日本最大級のAI向け計算性能を持つAI4Sスパコン（R7年度末までに整備）
- ✓ 富岳との接続によるシミュレーションとAI開発の高度な連携

共通基盤

実験・研究自動化
データ学習手法開発
AIエージェント開発

高橋恒一
(理研BDR)

＜理研の強み＞
- ✓ １０年前から先駆的に取り組んできた実験自動化の基礎技術開発による、ライフサイエンス分野の自動実験の技術基盤

※現在9センターが参画
（生命機能科学研究センター、バイオリソース研究センター、脳神経科学研究センター、生命医科学研究センター、創発物性科学研究センター、光量子工学研究センター、計算科学研究センター、革新知能統合研究センター、開拓研究所）

AI for Scienceのための計算基盤の整備

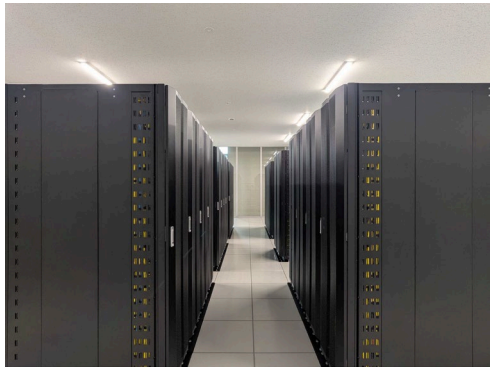
AI for Scienceにおいても、大規模言語モデル向け同等以上の計算能力が必要
例：ゲノム言語モデルEvo2のトレーニングでは、ABCI3.0数か月分を消費

理研計算科学研究センター（R-CCS）でのハードウェア・ソフトウェアの開発・整備

■AI for Scienceのための専用スパコンの導入
- ▷2024年度：パイロットシステムの導入
- ▷2025年度：エクサスケール級計算資源を導入
  - 最新GPUシステムNvidia GB200NVL4 x 400, 1600GPU
  - ABCI3.0の2/3程度の性能
- ▷**科学応用に必要な大規模推論計算**に注目した設計
  - 低精度の推論計算ではABCIを上回る性能
- ▷**AI for Science研究のための共用を進める計画**


松岡聡
(理研R-CCS)

■富岳でのシミュレーションとAIの融合

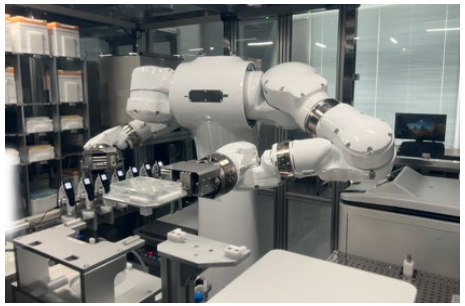


AIによる科学研究の自動化

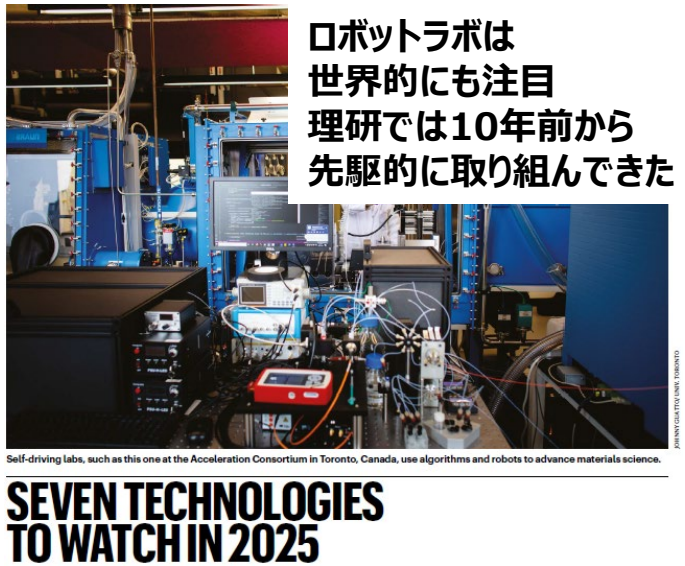
■AIとロボットで24時間365日実験を進める
未来型ラボ　→　**科学研究の自動化**

高橋恒一
(理研BDR)

自動実験系（ロボットラボ）の開発



Nature誌
2025年の注目技術7選
１．自律駆動ラボ
４．生命科学基盤モデル



ロボットラボは
世界的にも注目
理研では10年前から
先駆的に取り組んできた

事例紹介：Agentic AIのトラストの問題の研究動向

2026/2/19 EIP111にて発表したものの改訂版

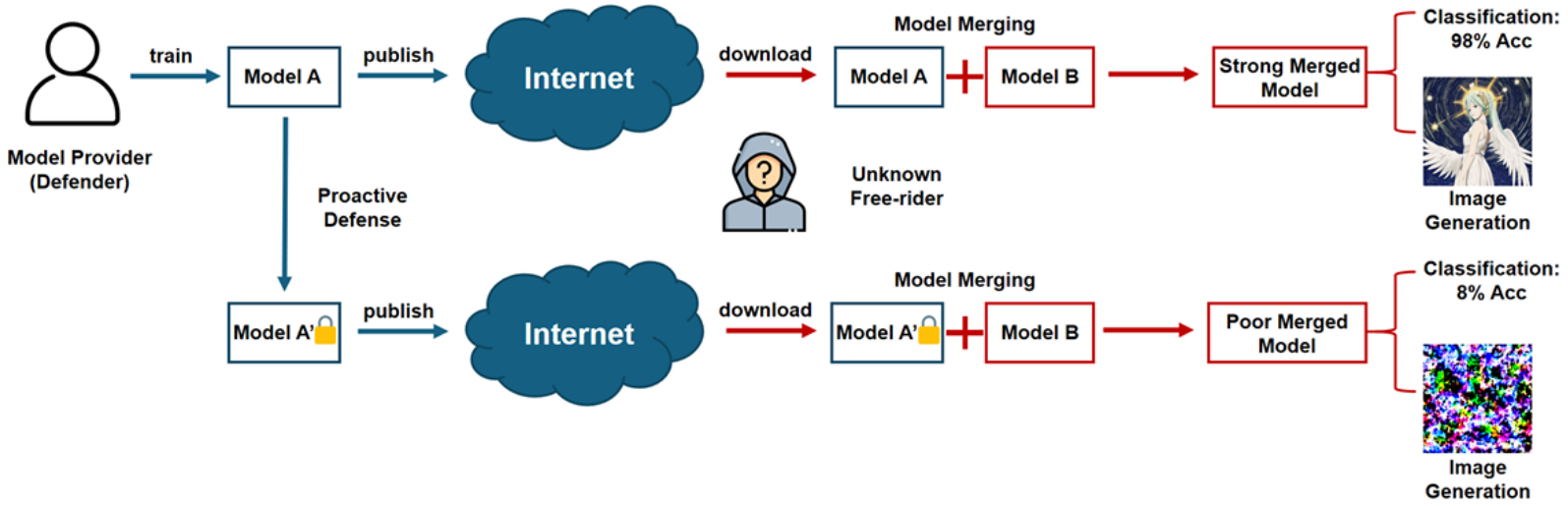
生成AIの重要な応用である Agentic AIのトラストの問題点に関する最近の研究動向

- これからはAgentic AIに関する研究が主流
- LLMが中核であり、LLMの問題はAgentic AIにも現れる。
- 予期される課題も未解決が多い。

中川裕志

理化学研究所・革新知能統合研究センター・チームディレクター
（任期満了により3月末で退職）

モデルマージに対するプロアクティブ防御 (ICCV 2025)



Research Question

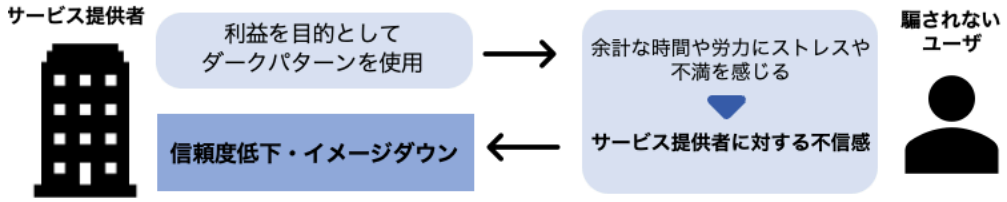
- モデルマージはそれぞれ異なるタスクに関する能力を持つ複数のopen weightモデルをトレーニングなしで一つのマルチタスクモデルに統合可能
- モデルマージを特定の者のみに許可し、フリーライドを防ぐことは可能か？

The Solution: PaRaMS

- モデルマージに対するプロアクティブ防御を実現初めての手法を提案
  - 秘密鍵を持つ者はモデルマージが可能
  - 秘密鍵を持たない物はモデルマージをするとモデル性能が大きく劣化
- 画像生成モデル、大規模言語モデルに適用可能

W Junhao, Y Zhe, J Sakuma, Disrupting model merging: A parameter-level defense without sacrificing accuracy, ICCV'25

ダークパターンの悪影響



最終的には**ダークパターンの利用が，**
企業の目的に反して利益を損なう結果につながる可能性

調査方法

タスク調査
- ダークパターンを含んだWebサイト上でタスク調査を実施

調査シナリオ
- Webサイト上でデリバリーサービスを利用し，6人分の料理を注文
- 料理の選択から注文完了までの一連の流れをWebサイト上で擬似的に体験

図2：作成したタスク調査用Webサイト

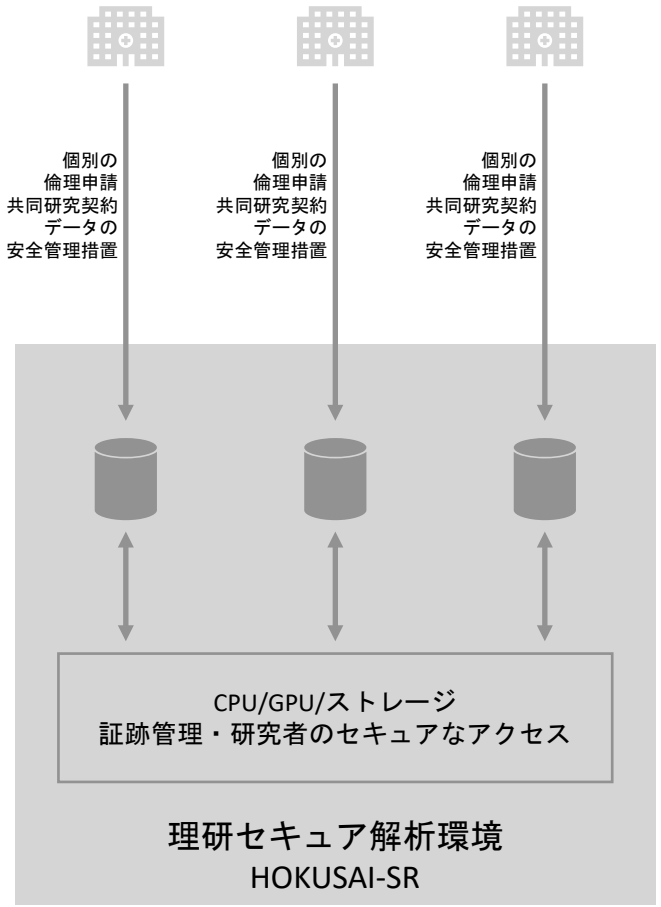
- 参加者からは、騙されたことへの不快感が表明された
- 購買意欲を低下させる可能性があることが示唆された
- ユーザーによっては、ダークパターンに気づいても、回避にかかるコストを見積もりながら、ダークパターンを受け入れることにしているケースもあり、ユーザー体験が悪化している可能性がある。

Kojima et.al., Are Dark Patterns Self-destructive for Service Providers?: Revealing Their Impacts on Usability and User Satisfaction, HCII 2024.

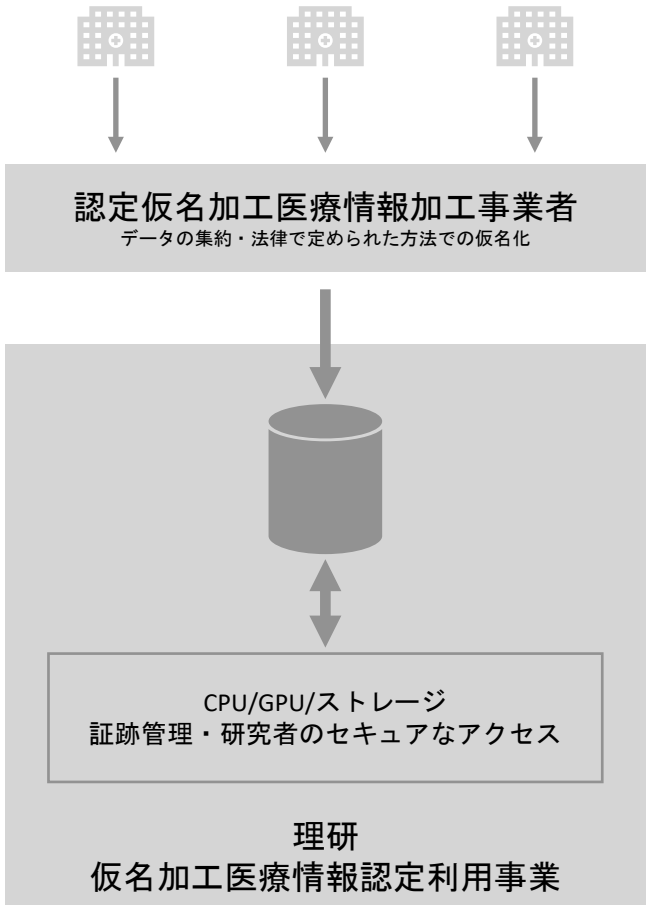
事例紹介：理化学研究所におけるセキュアなAI研究・開発用インフラの整備

医療AIなどの研究・開発を加速

従来の個人情報保護法



次世代医療基盤法



清田　純
情報統合本部・基盤研究開発部門
医科学データ共有開発ユニット　UL

まとめ

- 理研でのAIガバナンスについてのご紹介
- AI for Science関連やAI関連研究のご紹介
- 個別の事例のご紹介

- ありがとうございました。