CRDS 研究開発戦略センター
Center for Research and Development Strategy

国立研究開発法人科学技術振興機構

資料2－1
科学技術・学術審議会
基礎研究振興部会（第14回）
令和6年5月16日

文部科学省 科学技術・学術審議会 基礎研究振興部会 第14回

基盤モデル後のAI研究開発動向

2024年5月16日

科学技術振興機構(JST)
研究開発戦略センター(CRDS)

福島 俊一
toshikazu.fukushima@jst.go.jp
https://researchmap.jp/toshikazu_fukushima

基盤モデルに関する課題の全体観と次世代AIモデルの方向性

■ 日本国内で基盤モデル・生成AIの後追い開発や応用開発への取り組みが活発化している中、次世代AIモデルの創出につながる基礎研究の動向に着目

既に活発な取り組みが、国際的競争の中で進んでおり、走りながら迅速に手を打っていくべき課題

基盤モデル応用開発(API利用)
- チャットボット、仮想アシスタント、問い合わせ自動応答、質問応答
- コンテンツ生成(文章、画像、映像)
- 翻訳、要約、ライティング支援
- 企画支援、発想支援　他

分野固有基盤モデル開発・活用
- プログラミング向け基盤モデル
- 個別企業業務向け基盤モデル
- 法業務向け基盤モデル
- 医療・ヘルスケア向け基盤モデル
- 教育向け基盤モデル　　他

基盤モデル周辺拡張
- 基盤モデルが不得手な機能を扱う外部処理連携(最新情報検索、数式処理、物理シミュレーション、論理推論等)
- 問題解決ワークフロー設計の自動化
- プロンプトやワークフローの最適化　他

利活用時の問題対処の仕組み
- 生成AIの出力か否かを判別する仕組み(フェイク検出技術、電子透かしや識別情報付与の仕組みなど)
- 入出力データの著作権・肖像権関連問題への対処などのルール整備　他

基盤モデル運用
- 継続運用可能なビジネスモデル(ビジネス用途、研究用途)、エコシステム
- データ追加・更新プロセス
- トラストを確保した運営体制　他

基盤モデル構築
- 大規模深層学習モデル(トランスフォーマー、マルチモーダル)の実装
- 学習データの収集・選別・整備
- 大規模計算環境構築
- 高速化アルゴリズム、デバイス　他

次世代AIモデルの創出につながる基礎研究課題

AIリスク対処研究
- 基盤モデル自体の倫理性確保(RLHF等)
- 生成AI応用システムの品質管理(プロンプト型開発法のソフトウェア工学等)
- 人間・AI共生社会のリスク低減(エージェント設計論、トラスト形成等)　他

次世代AIモデル研究
- 基盤モデル高効率化、生成AI高性能化
- 基盤モデルのメカニズム解明
- 人間知能の理解に基づくモデルの探求、基盤モデルとの融合
- 新モデル向けコンピューティング　他

- 科学研究向け基盤モデル(AIロボット駆動科学)
- 大規模複雑問題解決

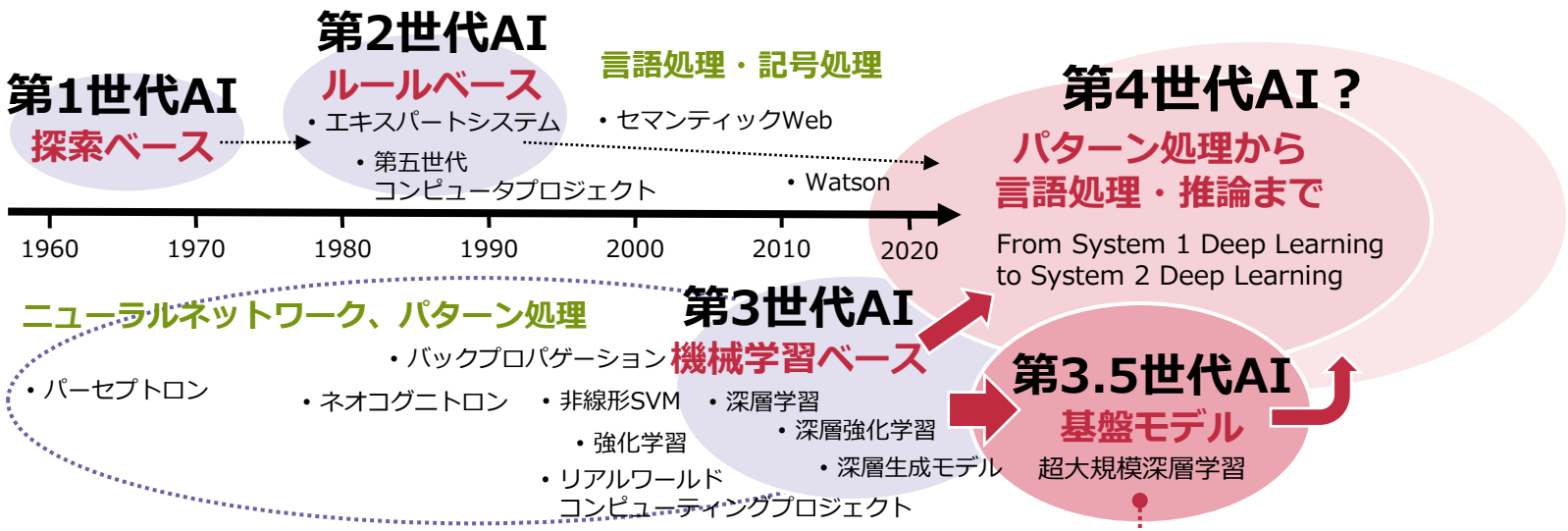
応用個別

共通基盤

実務 ←→ 学術

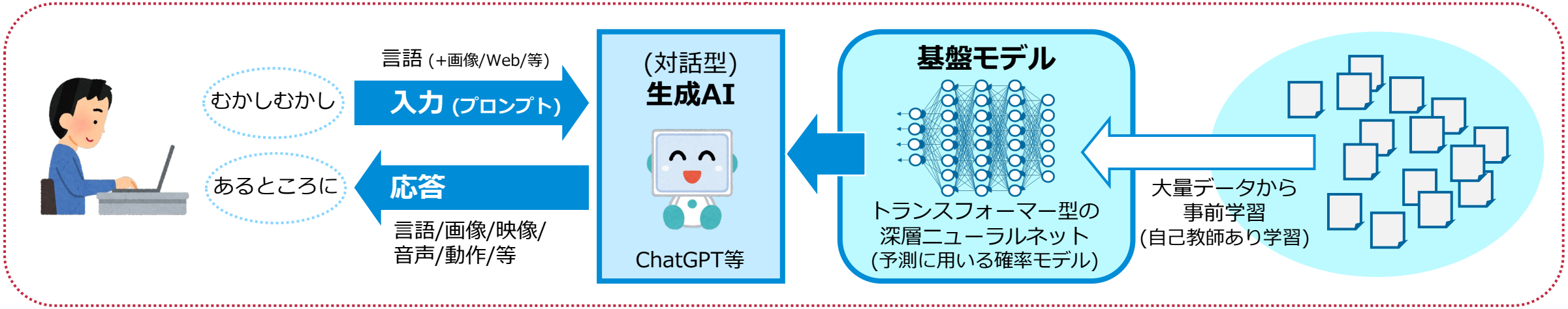
- 「人工知能研究の新潮流2　～基盤モデル・生成AIのインパクト～」(2023年7月) https://www.jst.go.jp/crds/report/CRDS-FY2023-RR-02.html
- 「戦略プロポーザル：次世代AIモデルの研究開発」(2024年3月) https://www.jst.go.jp/crds/report/CRDS-FY2023-SP-03.html

AIモデル発展の流れと現在の基盤モデル



現在の基盤モデル・生成AIは、超大規模学習で作られた確率モデルに基づく、いわば「高度なオートコンプリート機能」であることから来る問題・限界がある 次頁に記載

それらの問題・限界を克服するのが「次世代AIモデル」のチャレンジ



現在の基盤モデルの技術課題

■ それ以前のAIが目的特化型だったのに対して、高い汎用性を実現したことは画期的だが、
(人間の知能と比べると) 原理的に以下のような点に技術課題がある

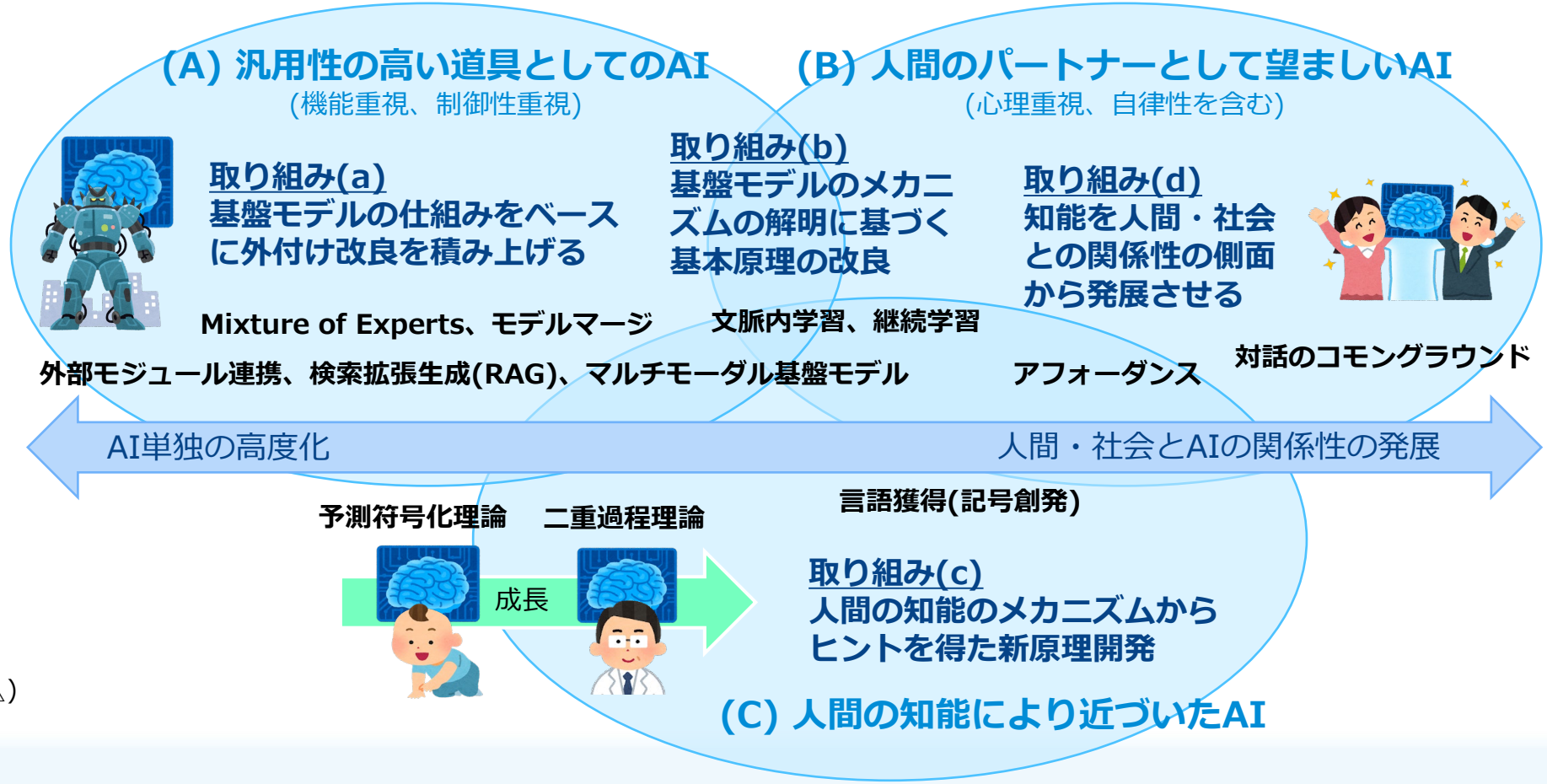
資源効率	極めて大規模なリソース(データ、計算機、電力等)が必要 • 最先端の基盤モデルは、1回の学習実行に数十億円の計算費用がかかる • 一方、人間の脳と消費電力は20ワット程度だといわれている
実世界操作(身体性)	動的・個別的な実世界状況に適応した操作・行動が苦手 • 基盤モデルの学習データは、いわば仮想世界での経験であり、実世界状況の動的変化や個別性に必ずしも適応できていない
論理性	論理構築・論理演算や大きなタスクのサブタスク分解が苦手 • 確率モデルに基づいて可能性の高い予測結果(応答)を返すものなので、論理構築や論理演算は行っていない
信頼性・安全性	人間と同じ価値観・目的を持って振る舞うと必ずしも信じられない • 基盤モデルはブラックボックスであり、どのような傾向・バイアスを持ったものか分からない • 確率モデルに基づくので、どうしても不安定性は残り、結果の100%保証はできない
自発性	行動の動機や目的を自ら生み出すことができない • 基盤モデルに限らず、現在のAIモデルは目的や価値基準は外部から与えるものである【これを求めるかは要議論】

課題克服に向けたアプローチ・研究動向

- 課題克服に向けた複数の取り組みが進みつつあるが、現状、課題解決の見込みは限定的または未知
- 目指すAIの姿として、**(A)汎用性の高い道具**、**(B)人間のパートナー**、**(C)人間の知能に近づく**、といった方向が見られるが、自律性が高まるにつれて、これらの取り組みは徐々に融合して発展

克服すべき問題点	取り組み(技術チャレンジ)			
	(a)	(b)	(c)	(d)
資源効率	×	?	○	?
実世界操作(身体性)	△	?	○	△
論理性	△	?	○	△
信頼性	△	?	△	○
安全性	△	?	○	○

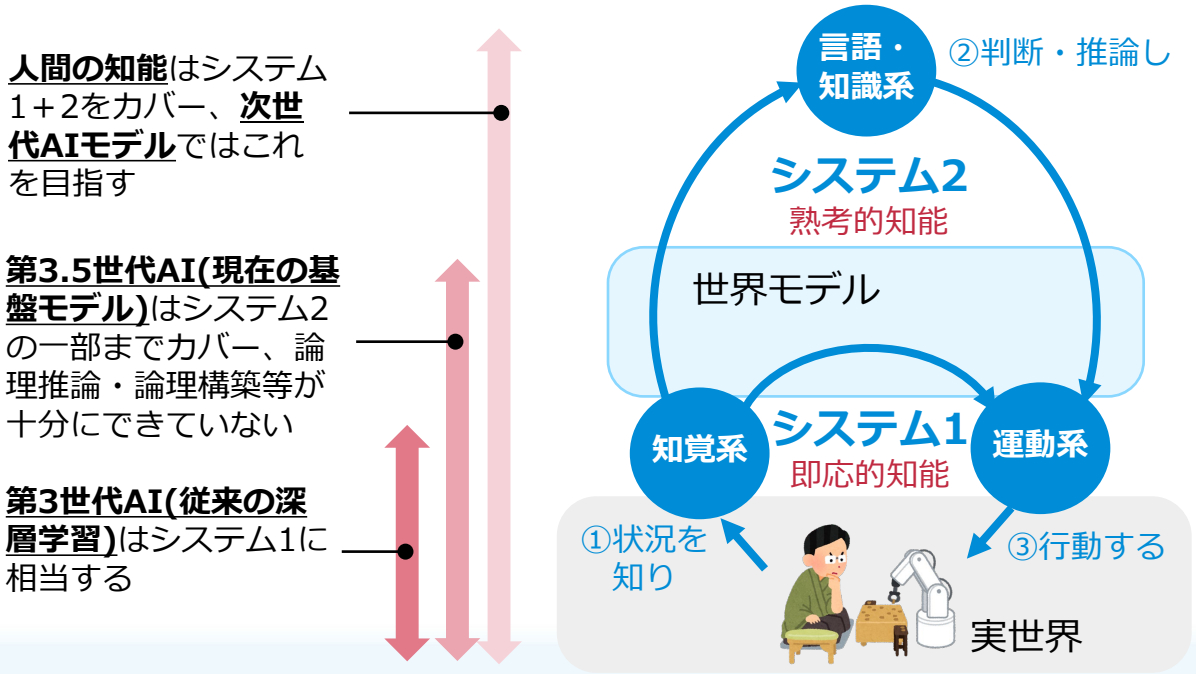
○ 効果が期待できる
△ 効果は限定的
× 悪化する
? 現時点では未知（○/△）



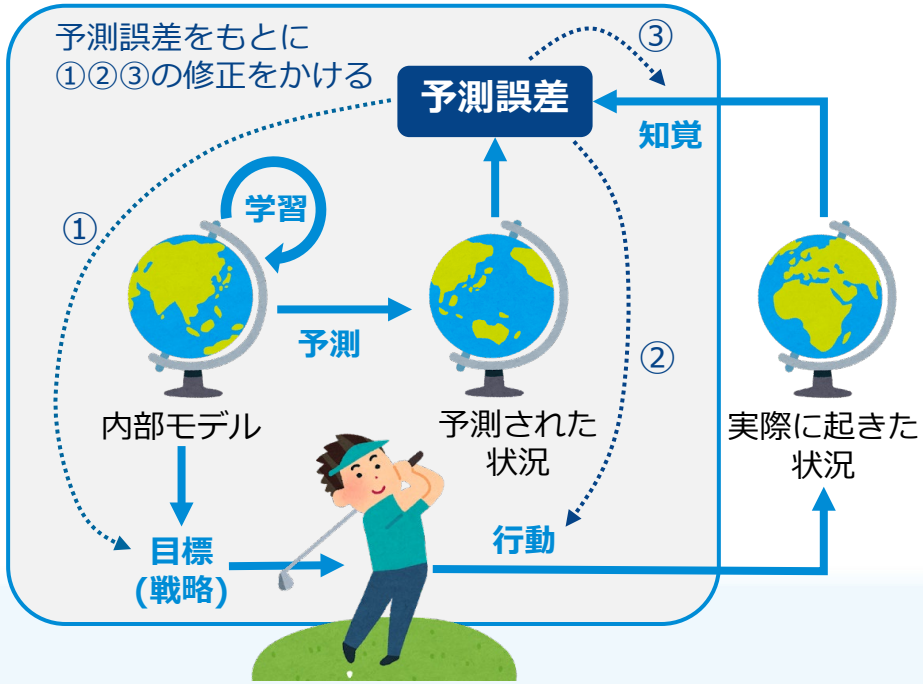
必ずしも大量事前学習を必要としないAIモデルの可能性

- 人間の知能自体は未解明だが、様々な知見が得られつつあり、現在のAIの課題克服につながり得る
- 現状の帰納型(ボトムアップ)AIでは膨大な事前学習を必要とするが、必要な範囲のみ能動的(トップダウン)に取りに行くような仕組みによって、資源効率を改善し得る
- 予測誤差からモデルを修正する仕組みは、大量の教師あり事前学習を必ずしも必要としない

二重過程理論：経験に基づいた即応的な情報処理を担うシステム1と、抽象化されたモデル・知識を参照した熟考的な情報処理を担うシステム2から成るという知能のモデル



予測符号化理論：乳幼児からの成長のように、他者や環境との相互作用を通じて、自己・環境の認知、言語獲得、行動・推論等の認知機能を発達させていく過程を、予測誤差最小化原理(自由エネルギー原理)によって統一的に説明



実世界操作の実現に向けたロボット基盤モデル開発

■ より多様でより大量のマルチモーダルデータを学習したロボット基盤モデルを構築し、
実世界におけるロボットの行動計画・動作生成を柔軟化・ロバスト化する試みが急速に進展

Google

① **PaLM-SayCan：自然言語による曖昧な要求**に対して、何ができるか、ロボットが行動を選択して実行

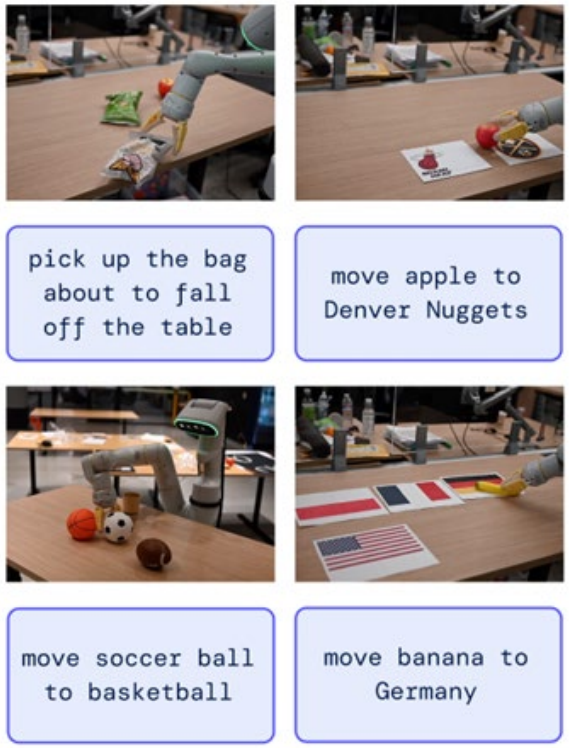
② **RT-1：実機を用いた大規模学習**によってロボット動作生成の汎用性を高めた

ロボット実機13台で17カ月、700以上のタスクをカバーする13万エピソードの動作データを収集・学習、700種類のタスクで97%の成功

③ **RT-2：Web上のテキストと画像も学習**することで、RT-1モデルで未学習だった物体も操作可能

④ **RT-X：**世界33研究機関が参加する**史上最大のオープンソースロボットデータセットプロジェクト**

RT-2において創発されたロボット動作生成例(一部)
https://doi.org/10.48550/arXiv.2307.15818



OpenAI+Figure

● **ChatGPT+Figure01：言語で対話、視覚や記憶から説明、計画して行動、推論して理由説明**等をデモ

「何か食べ物をもらえますか？」
「テーブルにある食べられるものはリンゴだけだったので、それをあなたに渡しました」

https://www.youtube.com/watch?v=Sq1QZB5baNw

①Google+Everyday Robots (2022年8月)、② Google+Everyday Robots (2022年12月)、
③Google DeepMind (2023年7月)、④Google DeepMindほか: Open X-Embodiment発表(2023年10月)、
その後、Google DeepMindは AutoRT、SARA-RT、RT-Trajectoryも発表 (2024年1月)

CRDS

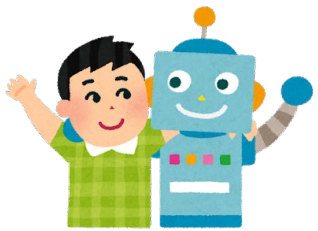
次世代AIモデルがもたらす社会的価値

■ 現在のAIが抱える、資源効率、実世界操作(身体性)、論理性、信頼性、安全性の課題が、克服・軽減されることで、例えば以下のようなことが可能になると期待

AIモデルの開発・更新のために、現在の基盤モデルのような膨大なデータ量・計算資源・電力消費は必要なくなり、環境負荷が低減される

生成AIの出力やその応用システムの振る舞いの正確性・倫理性・安全性が高まり、産業・教育・科学研究などさまざまな分野での活用が広がり、生産性が向上する

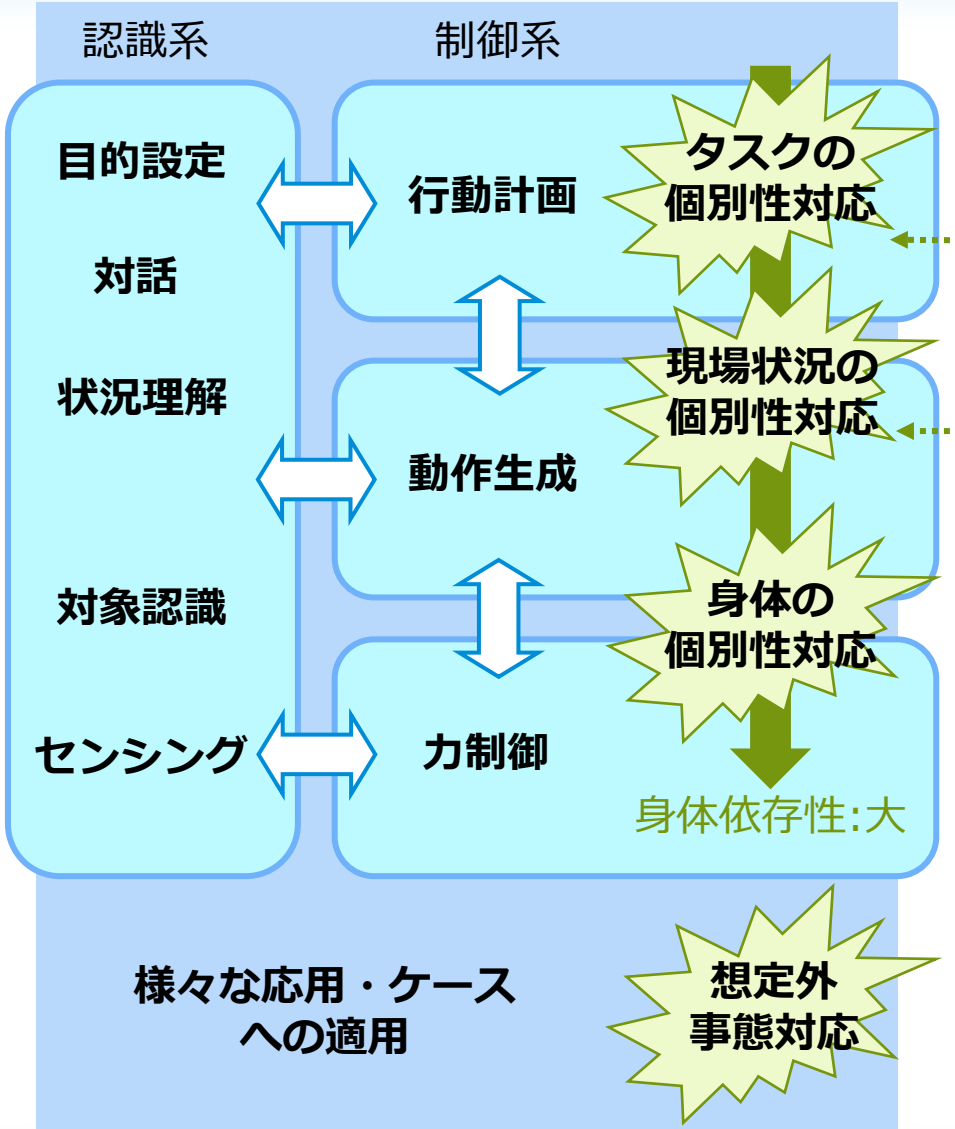
ハルシネーションの抑制やフェイクの判別が現在より進めやすくなり、不正確な情報や偽情報の流通による社会混乱や犯罪(詐欺・なりすましなど)の防止に役立つ



ロボット、ドローン、自動運転車などの動作・走行が、実世界の状況・場面に適応して柔軟に制御可能になり、より幅広い状況・場面での活用や安全性の向上につながる

AIが物理的な身体機能を獲得(AI×ロボット)すると、ホワイトカラーからブルーカラーまで幅広く労働市場に波及、大きな社会・経済的インパクトをもたらし得ることから、取り組みが急速に活発化している

AI×ロボットへの技術発展における課題と方向性



①次世代AIモデル

現在の基盤モデルの課題である資源効率・実世界操作(身体性)・論理性などを克服し、高い精度・効率と対話能力を持ち、タスクや現場状況や身体の個別性に柔軟に適応できるAIモデルを実現。また、学習から認知発達や進化への発展も探求。

②AIと身体機能システムの融合

①が非身体(AI)側からのアプローチであるのに対して、②は身体機能の側からのアプローチである。タスク/現場状況/身体の個別性に対して、ロボット側の個別対応(作り込み)を極力減らして柔軟に適応かつリアルタイムに動作できるようにする。さらに、所与の身体の制御に限らず、多様な身体への適応的制御や身体形状の適応まで可能性を探求。

③AIロボットと人の安全な協働・共生

機械学習ベースのAIモデルは、原理的に100%の精度保証・動作保証はできない。現在は事前統制されたClosed環境を用意することで安全性を確保しているが、Open環境では想定外の状況が起こる。応用に合わせてClosedからOpenへ段階的に環境を拡張しつつ、想定外リスクを可能な限り回避・対策し、AIロボットと人が安全に協働・共生できるようにする。

CRDS