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

国立研

資料６
科学技術・学術審議会　情報委員会
情報科学技術分野における戦略的重要
研究開発領域に関する検討会（第1回）
令和６年4月24日

文部科学省 科学技術・学術審議会 情報委員会
情報科学技術分野における戦略的重要研究開発領域に関する検討会（第1回）

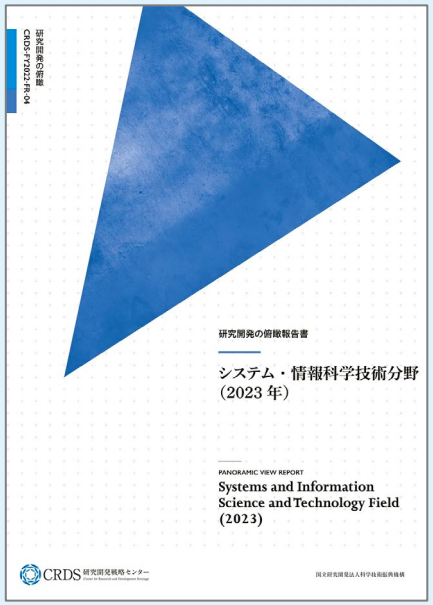
重点的に議論すべき領域検討に向けて

～CRDSでの俯瞰的調査と戦略提言から～

2024年4月24日

国立研究開発法人科学技術振興機構(JST)
研究開発戦略センター(CRDS)
フェロー　福島 俊一
toshikazu.fukushima@jst.go.jp
https://researchmap.jp/toshikazu_fukushima

- CRDSでは情報技術分野の研究開発動向の俯瞰的調査を継続実施
- 生成AIの急速な発展と社会インパクトを踏まえ、次世代AIの戦略提言



「研究開発の俯瞰報告書　システム・情報科学技術分野(2023年)」(2023年3月)

https://www.jst.go.jp/crds/report/CRDS-FY2022-FR-04.html

「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)

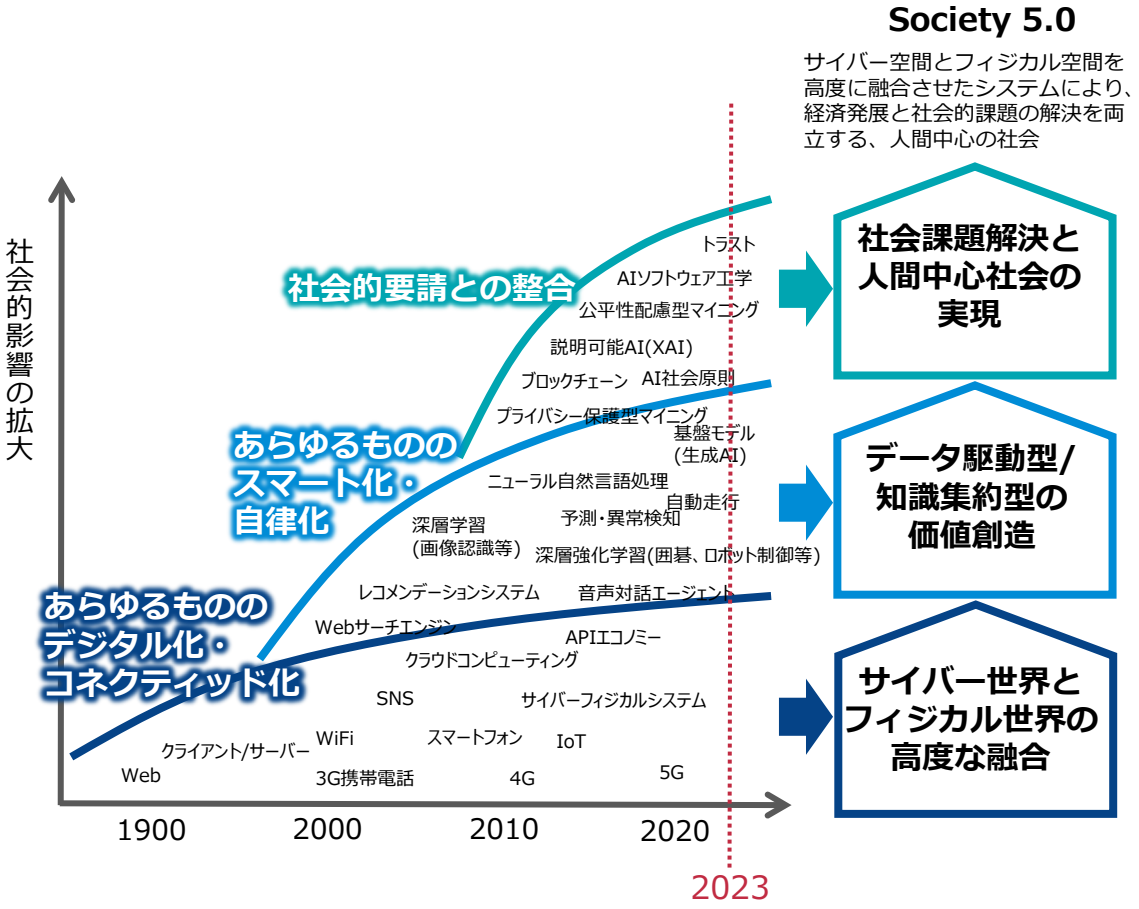
https://www.jst.go.jp/crds/report/CRDS-FY2023-RR-02.html

「戦略プロポーザル：次世代AIモデルの研究開発」(2024年3月)

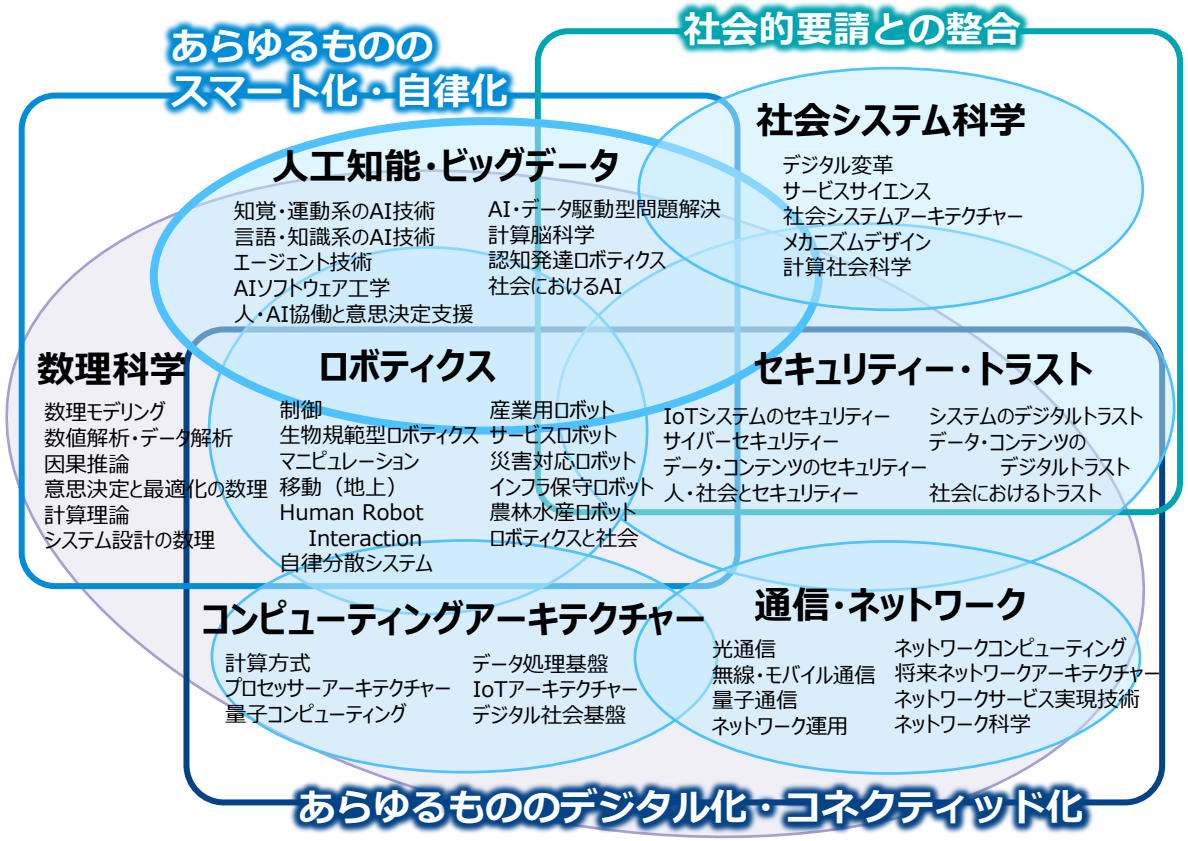
https://www.jst.go.jp/crds/report/CRDS-FY2023-SP-03.html

情報技術分野の俯瞰

情報技術分野のトレンドとビジョン

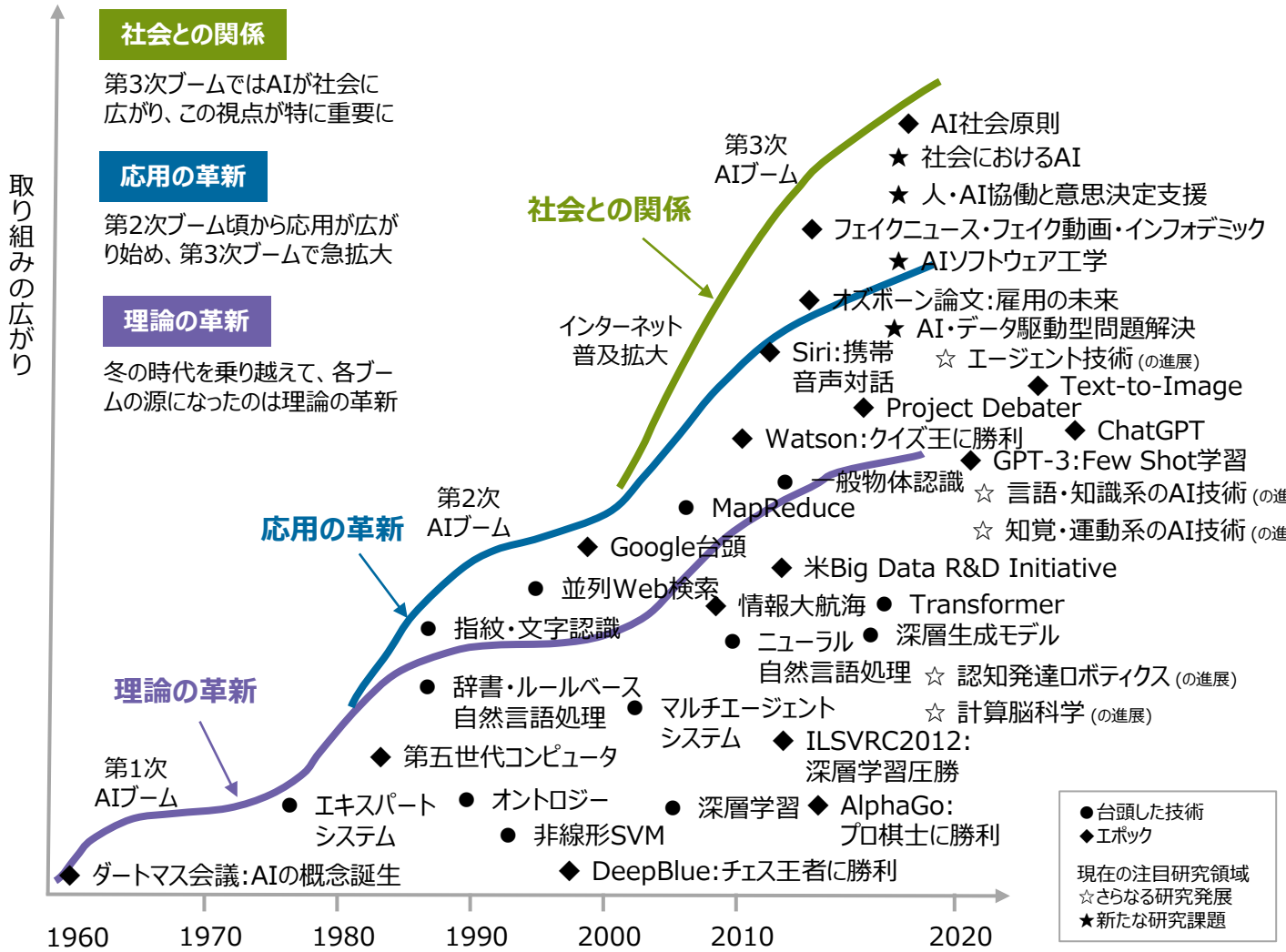


情報技術分野の俯瞰対象分野 (2023)



「研究開発の俯瞰報告書：システム・情報科学技術分野(2023年)」

AI・ビッグデータ分野の俯瞰「2つの潮流+1」



AI研究の2つの潮流+1 を軸に
9つの研究開発領域を設定して動向まとめ

Web・文献等の調査だけでなく、
有識者ヒアリング、ワークショップ等を重視
(個別インタビュー100名程度/年、学会・講演会参加100件以上/年)

「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～」

CRDS

次世代AIモデルの研究開発に関する戦略提言

■ 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推進強化を提言

基盤モデル・生成AIに関わる課題の全体観

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駆動DX
- 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

信頼されるAI
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

第4世代AI
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

応用個別 ③
共通基盤 ①②

実務 ⟷ 学術

本日の論点との対応

論点2
AI×ロボティクスやAI for Science等を含む、AIと他分野の連携領域

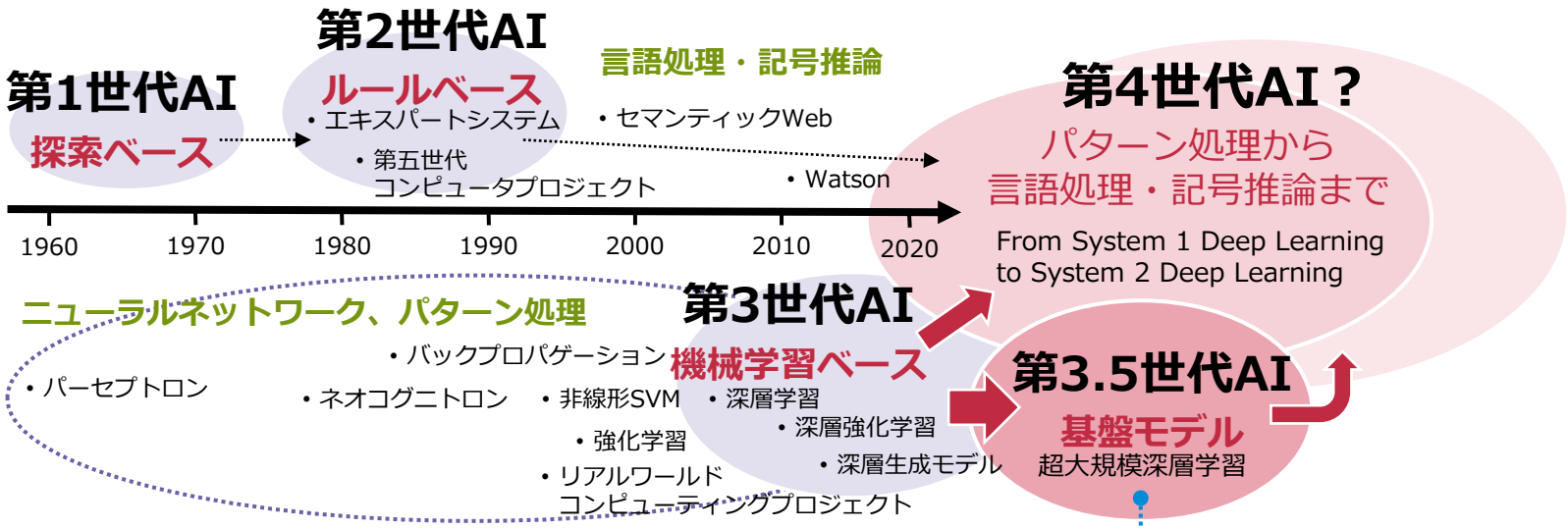
論点1
現在主流のAIの弱点等を克服するための革新的なAI関連技術等に関する研究領域

以降、①②③の注目動向を紹介

「戦略プロポーザル：次世代AIモデルの研究開発」

①-1　現在の基盤モデルの限界・問題点の克服へ

■ 現在の基盤モデル(生成AI)は、第3世代の深層学習を超大規模化した第3.5世代AIであり、驚異的な性能を示す一方で、帰納型の確率モデルであることによる限界・問題点が存在する



限界克服へのアプローチ

a. 現在の基盤モデルを出発点とした改良・発展の研究
b. 人間(生物)の知能からヒントを得た新原理研究
c. 他者・環境との関係性の中で発展する知能の研究

※アプローチを限定せず幅広に可能性探索、ただし、異なるアプローチ間で知見共有しシナジーを生み出す

現在の基盤モデルの限界・問題点

1. **資源効率：**極めて大規模なリソース(データ、計算機、電力など)が必要
2. **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
3. **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
4. **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

①-2　a. 現在の基盤モデルを出発点とした改良・発展の研究

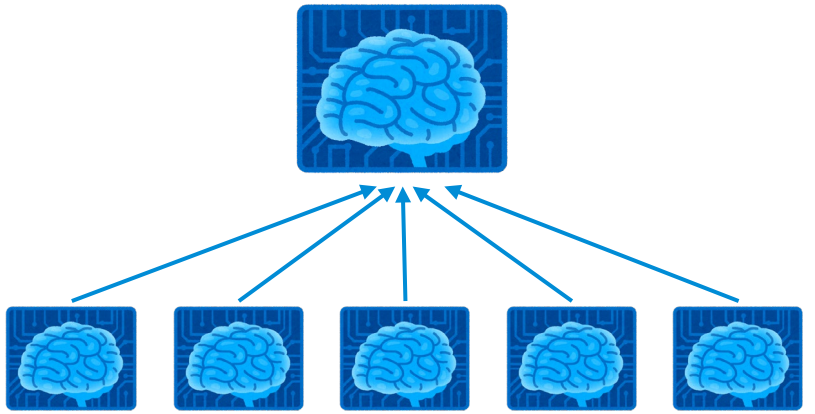
- 現在の基盤モデルでなぜあれほど賢い振る舞いをするのか分かっておらず、そのメカニズムを数理的に解明する研究は、より高い性能でより効率の良いAIモデルを生み出すための基礎として重要
- 現在の基盤モデルが苦手な問題に対処するための仕組みを外付けで実現する取り組みは既に多数実現されている　(ただし、資源効率問題への根本的対処は難しい)
  - プラグイン、 RAG(検索拡張生成)、アンサンブル的手法(Mixture of Expertsやモデルマージ) ほか

LLM-jp

- 国立情報学研究所(NII)を中心に国内の自然言語処理・計算機システムの研究者が多数参加
- コーパス構築WG、チューニング・評価WG、mdxWG、モデル構築WGなどが立ち上がり、研究成果・知見、データ・計算資源を共有しつつ、大規模言語モデル(LLM)のメカニズム理解と研究用オープンソース日本語LLM開発を推進
- 2024年4月にNIIにLLM研究開発センター設置

Mixture of Experts

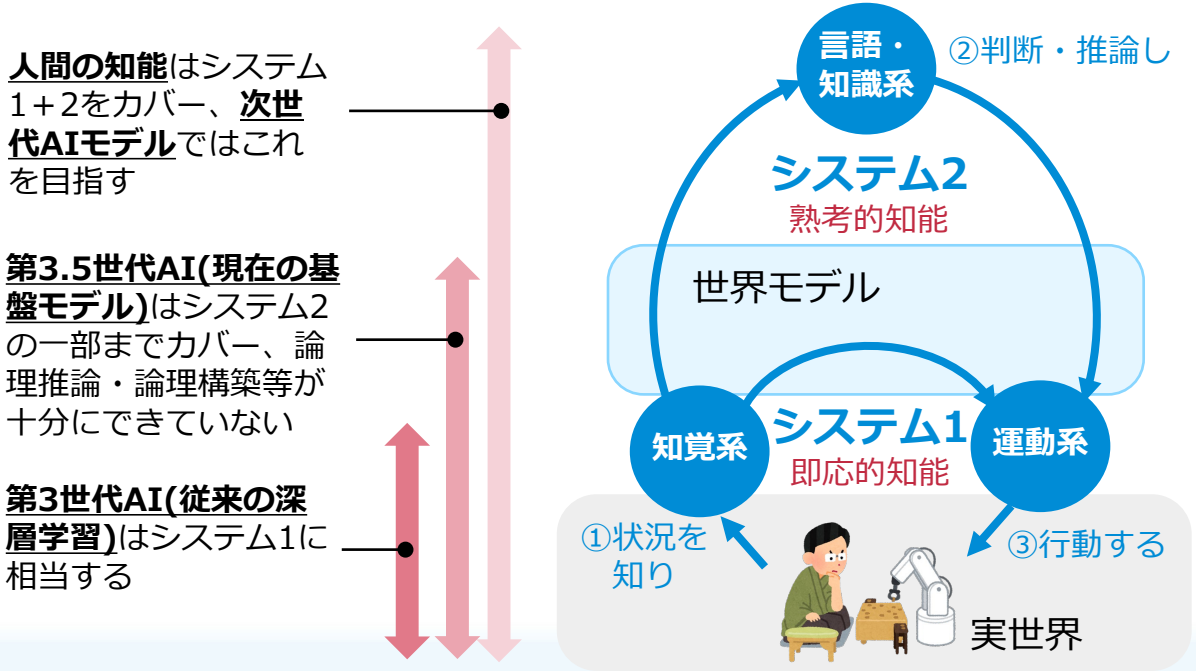
異なる得意・特性を持ったエキスパートモデルから結果を集めて総合的に判定



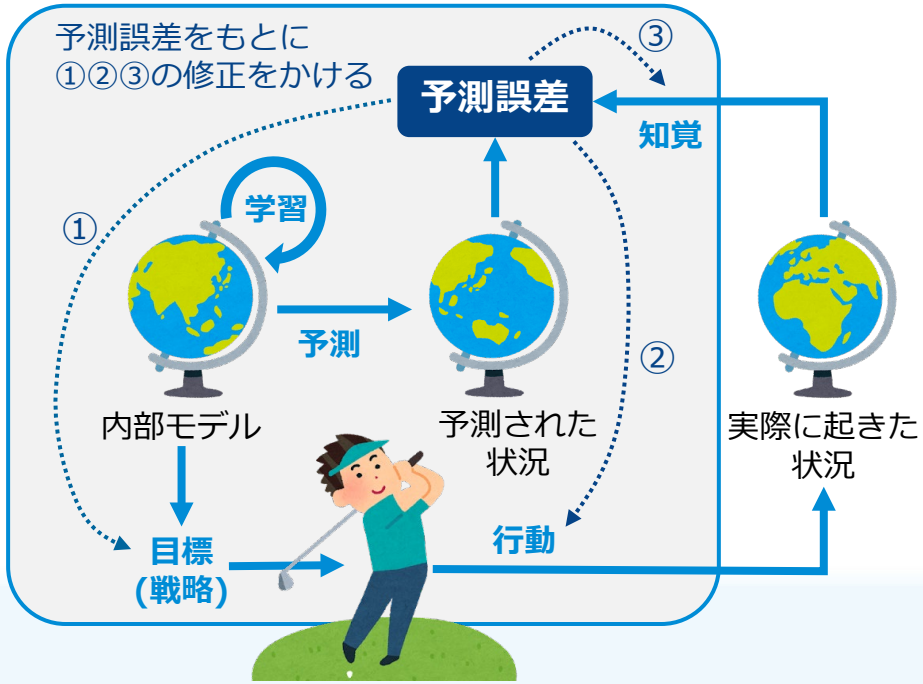
①-3　b. 人間(生物)の知能からヒントを得た新原理研究

- 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

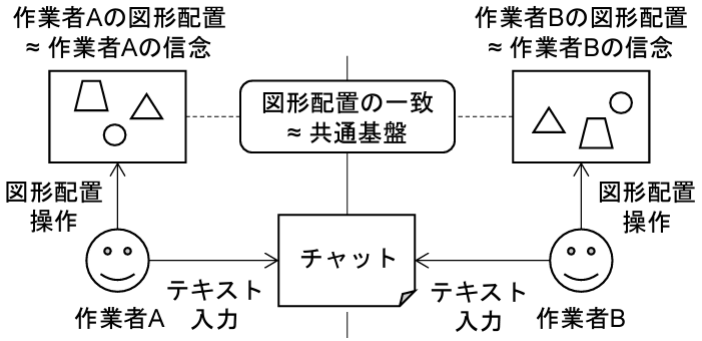


①-4　c. 他者や環境との関係性の中で発展する知能の研究

- アプローチa・bは、ひとつのAIシステムをより高度化させる取り組みだが、アプローチcは、他者や環境との関係性の中で知能が発展すると考える
- 現状、必ずしもまとまった大きな流れになっているわけではないが、多数のAIと人間が協働・共存するマルチエージェント社会において考えておくべきAIモデルの観点がいくつか指摘されつつある

コモングラウンド

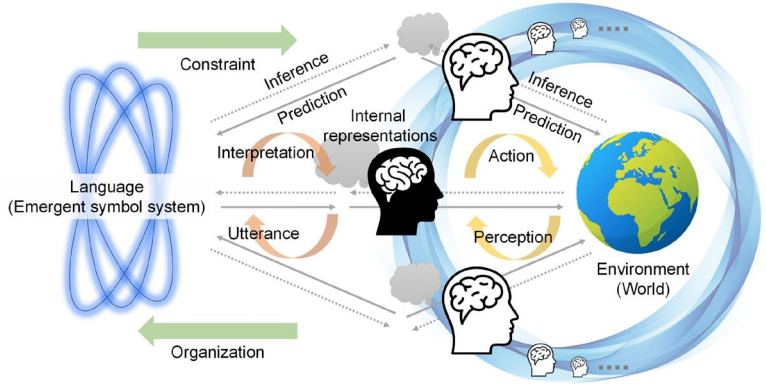
コミュニケーションを取る上で欠かせない、相手との共通理解や会話のバックグラウンドを「コモングラウンド」という。2者間でコモングラウンドを持つことで、共同作業による目的達成が可能になる。現在のChatGPTと利用者との間にはコモングラウンドがない。



自然言語処理 (2023)　https://doi.org/10.5715/jnlp.30907

言語獲得(記号創発)

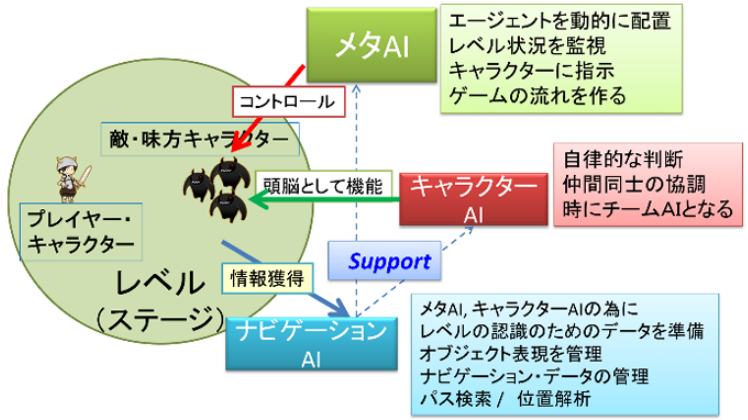
どのような概念を切り出し、それにどのようなラベル(言葉)を与えるかは、集団の中でのやり取りを通して創発的に決まっていく。科学的知識の形成もこれと類似したものかもしれない。このような創発のメカニズムが検討されている。



認知科学 (2024)　https://doi.org/10.11225/cs.2023.064

アフォーダンス

環境や物体の側が、人間や動物に対して意味や価値を与えているという考え方である。コンピューターゲーム(仮想空間)の中にAIを組み込む事例などで、この考え方を用いて効率を高めることなどが行われている。



人工知能学会論文誌 (2020)　https://doi.org/10.1527/tjsai.B-J64

②-1　生成AIによって深刻化するAIリスク

■ 基盤モデル・生成AIは、それ以前の目的特化型AIを超えて高い汎用性とマルチモーダル性を持ち、人間と区別できないような応答性能を示すことから、様々な面でリスクの深刻化を招いた

生成AIの出力から生じる問題

- ウソや架空の出来事をあたかも事実であるかのように語る(**ハルシネーション**)

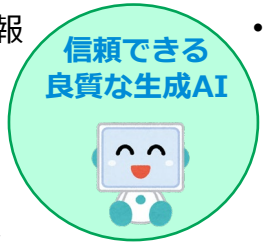
- 差別・偏見、偏った価値観が応答中に表れる(**社会的バイアス**)
- 学習データやプロンプトから**個人情報・機密情報が漏洩**

- 学習データや生成データの**著作権・肖像権の問題**
- **クリエーターや俳優・声優などの反発**、創作市場・文化への影響

- AIによる**労働者の置き換え・失業**
- 学習過程における**低賃金労働者搾取**
- 大量の電力・水の消費による**環境インパクト**

生成AIのトラスト問題

- 個人情報・機密情報は**学習に使わない**



- **正確性・安全性・倫理等を確保**するようにモデルを調整

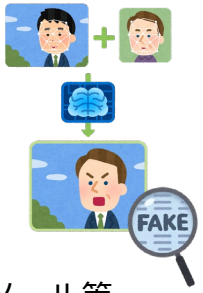
- 個人情報・機密情報の除外や**品質確保**をしていない
- **偽ってもバレない**だろう
- 犯罪向けの**ワームGPT**
- 特定主義・思想の**プロパガンダ**
- 利用者の**個人情報抜き取り**
- **汚染**されたモデル、仕込まれた**バックドア**

社会の在り方・文化への影響

- 超知能などが**予期せぬ挙動・事態**を引き起こす懸念
- **教育の在り方**への影響、AI依存による思考力低下の懸念

生成AIの悪用問題

- **フェイク動画**や**フェイクニュース**を生成、SNSで拡散して世論を誘導・干渉、ハラスメント・攻撃



- なりすましや詐欺メール等を生成し、**犯罪利用**
- 人々を思い通りに**誘導・洗脳**、思考停止させ、依存させる
- 武器や毒薬の作り方などの**悪知恵を聞き出し**(**脱獄**/Jailbreak)※

※生成AIに対するPrompt Injection攻撃の一種

- フェイク拡散による世論誘導・選挙干渉、対立激化による**民主主義の質的低下**
- 証拠の信憑性の低下による**犯罪捜査・司法のゆらぎ**

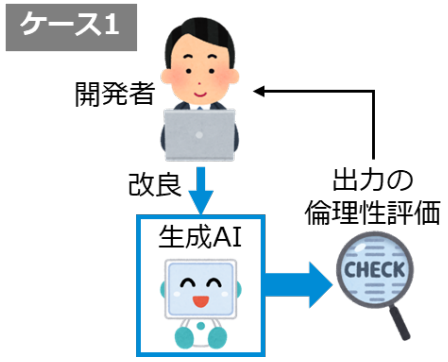
②-2　生成AIリスクへの対処技術開発

■ 次世代AIモデル研究では、高い性能・機能の追求だけでなく、リスクへの対処にも同時に取り組むべきであり、これには制度設計だけでなく、新たな技術開発が不可欠である

リスク対処が必要になる5つのケースと技術開発状況

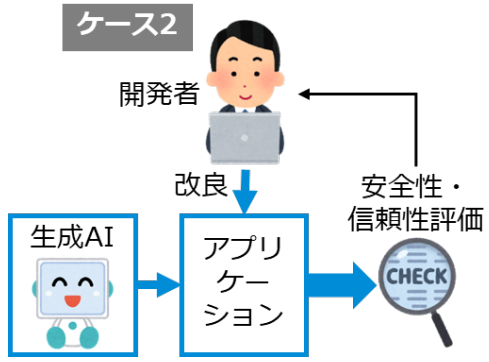
開発時

生成AI出力の倫理性確保（不適切な応答回避）

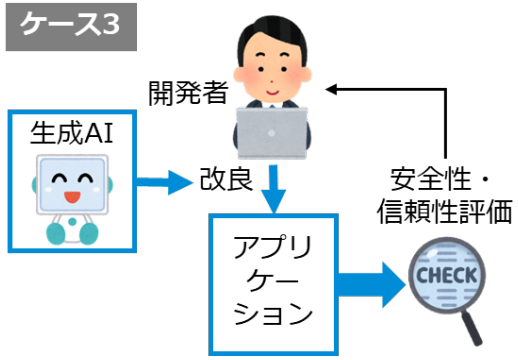


学習データ選別、不適切出力フィルタリングのほか、人間フィードバックによる基盤モデルチューニング(RLHF, DPO)等、活発に技術開発が進められている。

生成AIのアプリ開発での品質管理



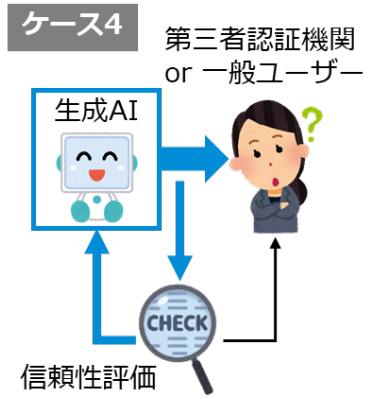
生成AIを利用して開発されたアプリの品質管理



2018年頃に機械学習とソフトウェア工学の融合分野(機械学習工学/AIソフトウェア工学)が立ち上がり、設計法、テスト法、説明性・公平性等の技術開発が進んでいる。AI応用システム品質管理ガイドライン(QA4AI, AIQM等)やAI事業者ガイドラインに、生成AI対応を盛り込む検討が進められている。ケース2対応は従来から検討されていたが、ケース3は生成AIで新たに必要になった。

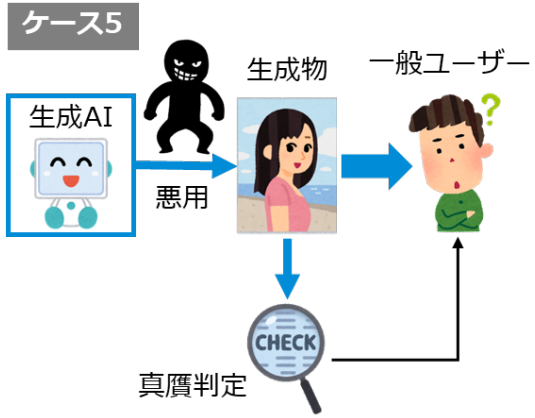
利用時

乱立した生成AIが良質なものかの審査等



生成AI乱立傾向に向かいつつあり、第三者認証機関の必要性が高まっている。どのような技術的検証手段が取り得るかは今後。

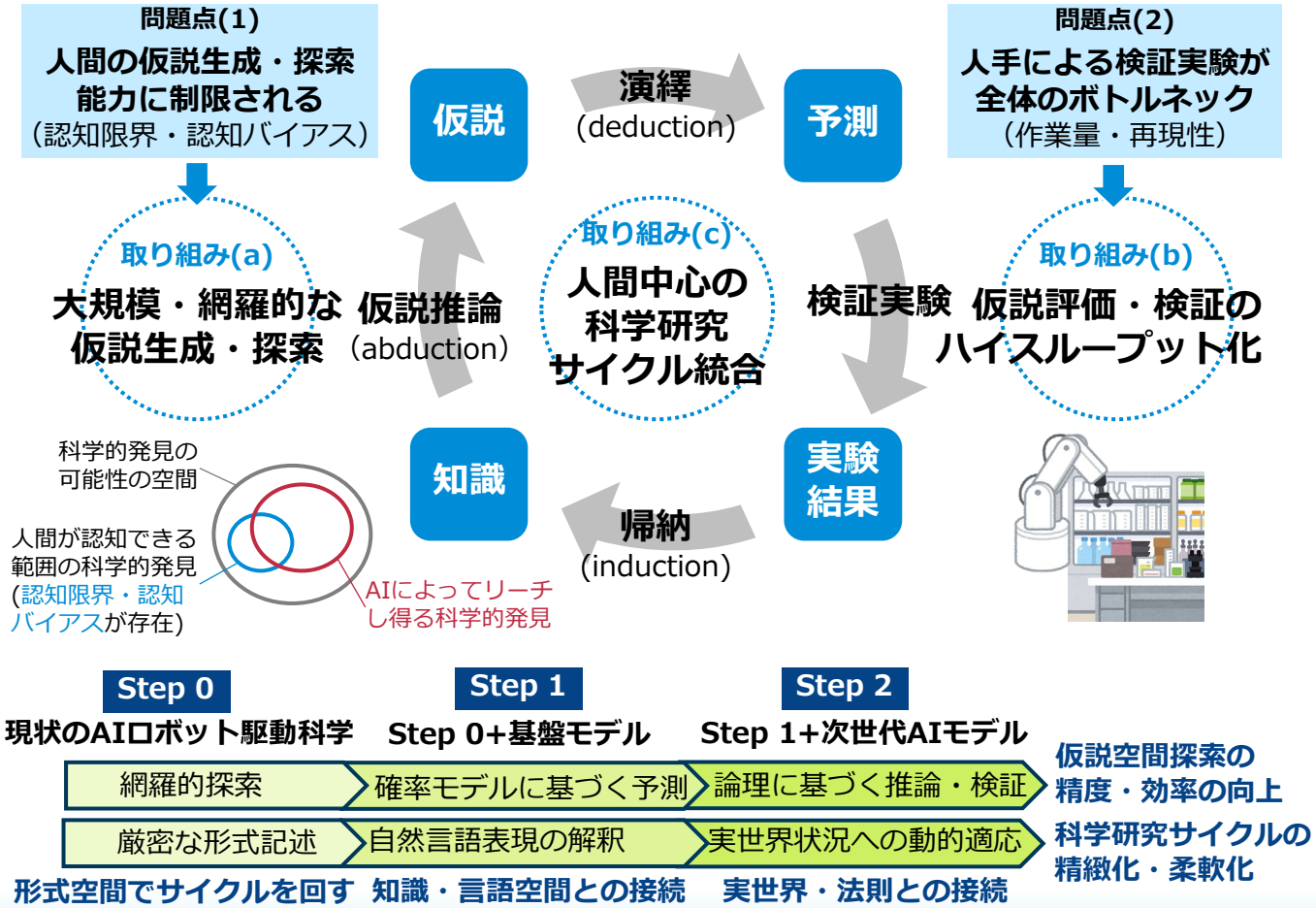
フェイクにだまされないようにする対策等



喫緊の課題として取り組み強化。オリジネータープロファイル等のアイデンティティ管理による出所・経路追跡と、コンテンツ詳細解析によるフェイク判別技術が取り組みの中心。

③　AI×科学、AI×ロボティクス

AI×科学：AIロボット駆動科学の枠組みに基盤モデルや次世代AIモデルが統合されることで、仮説空間探索の精度・効率が向上し、科学研究サイクルの精緻化・柔軟化が大きく進展する



AI×ロボティクス：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用いることで、行動計画・動作生成を柔軟化・ロバスト化

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**

⑤ **AutoRT：**基盤モデルとRT-1,2を組み合わせて多様な学習データ収集

⑥ **SARA-RT：**RTモデルの効率を改善

⑦ **RT-Trajectory：**学習動画にロボットの動きを説明する視覚的なアウトラインを追加、性能が2倍以上向上

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)、⑤⑥⑦Google DeepMind (2024年1月)

CRDS　「戦略プロポーザル：次世代AIモデルの研究開発」

【参考】AI研究開発の形態変化を踏まえたAI研究推進策

ビッグサイエンス化
基礎研究に必要な計算資源・データの超大規模化、その運営は研究者だけでなく多様なスタッフを擁した大規模プロマネが必要

ハイスピード化・ハイインパクト化
社会・生活に広く影響を与え得る技術、その影響範囲の予測困難な技術が次々に(数週間単位で)生まれている

非オープン化
ビッグテック企業が最先端の研究成果・知見を保有し、その内容が公開されない(公共財化されない)傾向が強まっている

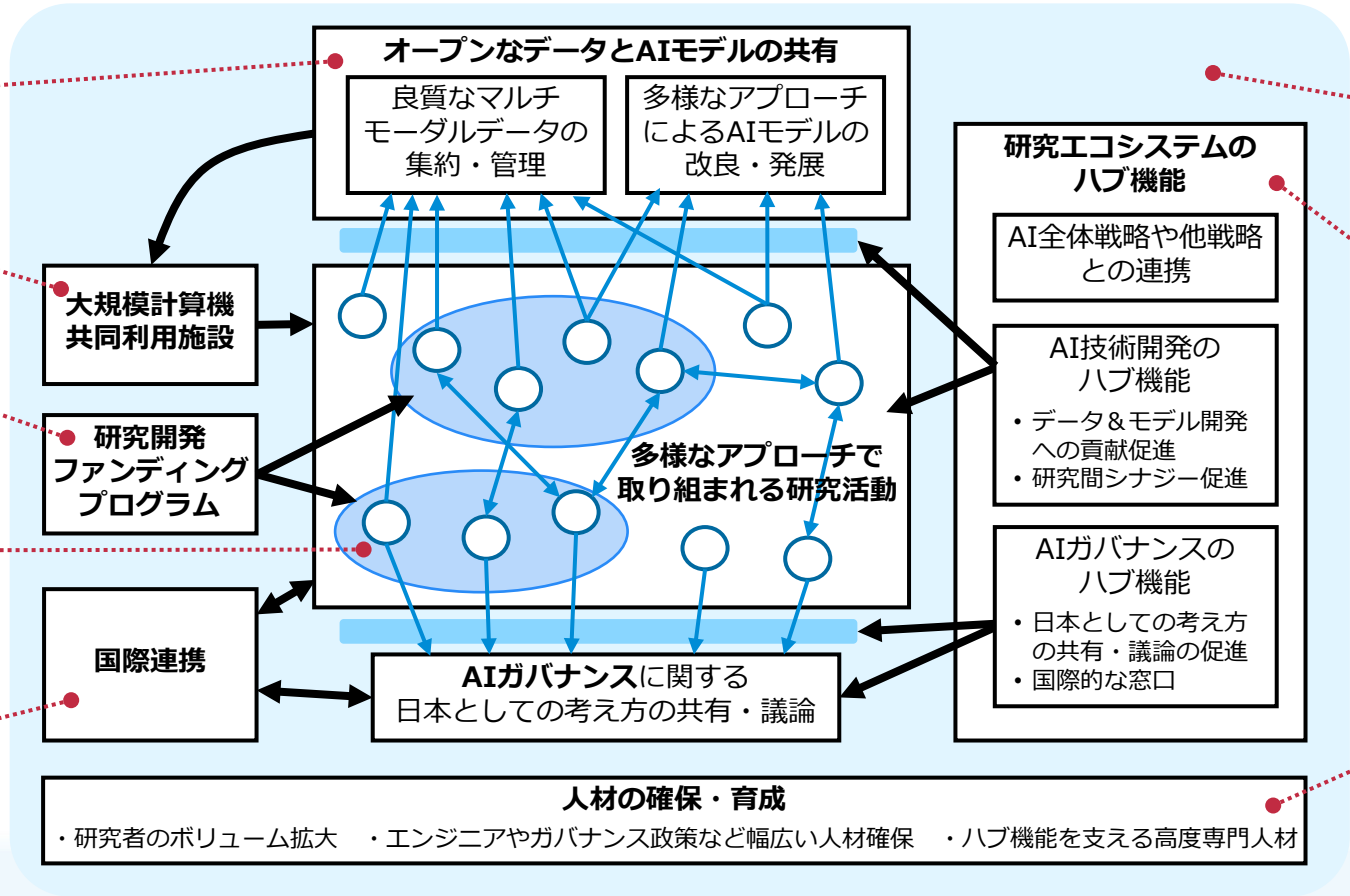
c.AIモデルとマルチモーダルデータの集約・共有・評価・管理体制の整備

b.大規模計算機の共同利用施設の継続的な運用・強化

g.研究エコシステムを生かした柔軟でアジャイルなプログラム運営

e.技術系研究者のみならず人文・社会系研究者の主体的参画を促進するプログラム設計

d.基礎研究とルールメイキングにおけるオープンな国際連携とその支援体制構築



a.さまざまな研究機関・組織が結集し協力し合う研究エコシステムの形成

f.研究エコシステムのハブ機能を担う組織の設置

h.研究エコシステムを支える人材の確保・育成

「戦略プロポーザル：次世代AIモデルの研究開発」