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３
情報委員会（第35回）
令和６年１月25日

人工知能(AI)研究の潮流

JST CRDSでの俯瞰的調査から

2024年1月25日

科学技術振興機構(JST)
研究開発戦略センター(CRDS)

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

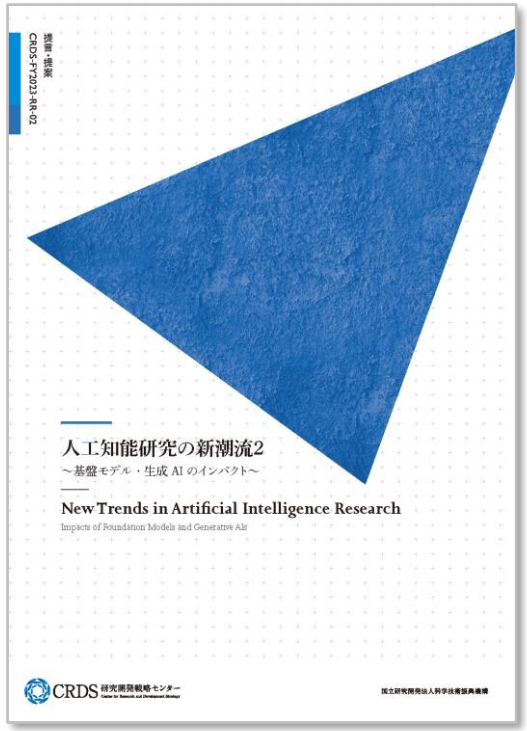
本発表に対応するCRDS報告書

研究開発の俯瞰報告書
システム・情報科学技術分野 (2023年)

CRDS-FY2022-FR-04 (2023年3月)

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

2.1節「人工知能・ビッグデータ」



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-FY2023-PR-02 (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(仮)

参考: JST CRDSの報告書　AI関連の既発行分

分野俯瞰

- 人工知能研究の潮流2 ～基盤モデル・生成AIのインパクト～ (2023年)
  https://www.jst.go.jp/crds/report/CRDS-FY2023-RR-02.html
- 人工知能研究の新潮流 ～日本の勝ち筋～ (2021年)
  https://www.jst.go.jp/crds/report/CRDS-FY2021-RR-01.html
- 俯瞰ワークショップ報告書：エージェント技術 (2022年)
  https://www.jst.go.jp/crds/report/CRDS-FY2021-WR-11.html
- 俯瞰ワークショップ報告書：ヒューマンインタフェース研究動向 (2023年)
  https://www.jst.go.jp/crds/report/CRDS-FY2022-WR-10.html
- 研究開発の俯瞰報告書：システム・情報科学技術分野 (2023年)
  https://www.jst.go.jp/crds/report/CRDS-FY2022-FR-04.html

戦略提言(1) AIソフトウェア工学

- 戦略プロポーザル：AI応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立 (2018年)
  https://www.jst.go.jp/crds/report/CRDS-FY2018-SP-03.html
- 科学技術未来戦略ワークショップ報告書：機械学習型システム開発へのパラダイム転換 (2018年)
  https://www.jst.go.jp/crds/report/CRDS-FY2017-WR-11.html

戦略提言(2) 意思決定・合意形成支援

- 戦略プロポーザル: 複雑社会における意思決定・合意形成を支える情報科学技術 (2018年)
  https://www.jst.go.jp/crds/report/CRDS-FY2017-SP-03.html
- 科学技術未来戦略ワークショップ報告書：複雑社会における意思決定・合意形成を支える情報科学技術 (2017年)
  https://www.jst.go.jp/crds/report/CRDS-FY2017-WR-05.html
- 公開ワークショップ報告書：意思決定のための情報科学～情報氾濫・フェイク・分断に立ち向かうことは可能か～ (2020年)
  https://www.jst.go.jp/crds/report/CRDS-FY2019-WR-02.html

戦略提言(3) 第4世代AI

- 戦略プロポーザル：第4世代AIの研究開発―深層学習と知識・記号推論の融合― (2020年)
  https://www.jst.go.jp/crds/report/CRDS-FY2019-SP-08.html
- 科学技術未来戦略ワークショップ報告書：深層学習と知識・記号推論の融合によるAI基盤技術の発展 (2020年)
  https://www.jst.go.jp/crds/report/CRDS-FY2019-WR-08.html
- JSAI2020企画セッション報告書：次世代AI研究開発―さらなる進化に向けて― (2020年)
  https://www.jst.go.jp/crds/report/CRDS-FY2020-XR-02.html

戦略提言(4) AI駆動科学

- 戦略プロポーザル：人工知能と科学 ～AI・データ駆動科学による発見と理解～ (2021年)
  https://www.jst.go.jp/crds/report/CRDS-FY2021-SP-03.html
- 俯瞰セミナーシリーズ報告書：機械学習と科学 (2021年)
  https://www.jst.go.jp/crds/report/CRDS-FY2020-WR-13.html
- 科学技術未来戦略ワークショップ報告書：人工知能と科学 (2021年)
  https://www.jst.go.jp/crds/report/CRDS-FY2021-WR-01.html
- 計測横断チーム調査報告書　計測の俯瞰と新潮流 (2018年)
  https://www.jst.go.jp/crds/report/CRDS-FY2018-RR-03.html

戦略提言(5) デジタル社会のトラスト

- 戦略プロポーザル：デジタル社会における新たなトラスト形成 (2022年)
  https://www.jst.go.jp/crds/report/CRDS-FY2022-SP-03.html
- 俯瞰セミナー＆ワークショップ報告書：トラスト研究の潮流
  ～人文・社会科学から人工知能、医療まで～ (2022年)
  https://www.jst.go.jp/crds/report/CRDS-FY2021-WR-05.html
- 科学技術未来戦略ワークショップ報告書：トラスト研究戦略
  ～デジタル社会における新たなトラスト形成～ (2022年)
  https://www.jst.go.jp/crds/report/CRDS-FY2022-WR-05.html
- 公開シンポジウム報告書「デジタル社会における新たなトラスト形成
  ～総合知による取り組みへ～」(2023年)
  https://www.jst.go.jp/crds/report/CRDS-FY2022-WR-05.html

**いずれも全文pdfダウンロード可能、
希望に応じて冊子版の送付も可能**

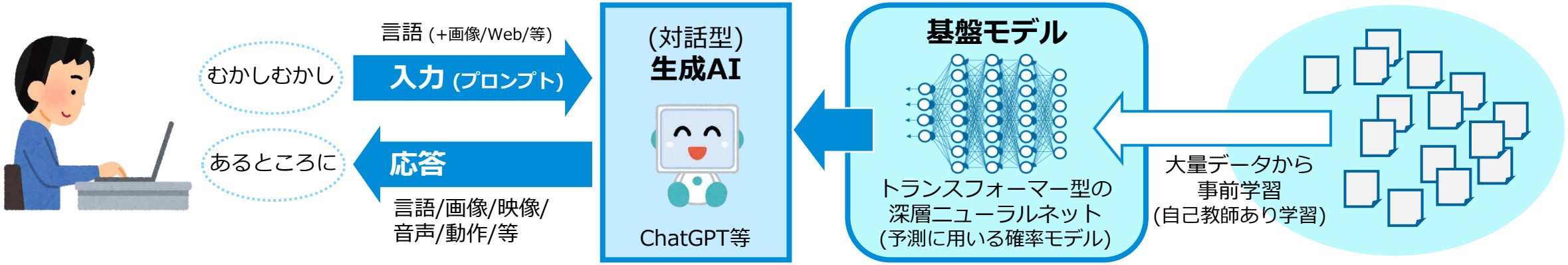
CRDS

目次

目次

基盤モデル・生成AIのインパクト

- 爆発的にユーザー拡大　ChatGPTは2022年11月30日公開後2ヶ月でアクティブ利用者数が1億人に到達
- 人間のような自然な応答、専門的な知識・能力を備えているかのような応答
- それまでの目的特化AIから汎用性・マルチモーダル性の高いAIへと発展
- 人間の知的作業全般を変革、産業、研究開発、教育、創作など様々な分野に幅広く波及



生成AIのユースケース例

自然言語処理	コンピュータビジョン	ソフトウェアエンジニアリング	インタフェース	科学・教育・医療等の応用
テキスト生成、質問応答、要約、検索、分類、意図認識、翻訳、リライト、音声テキスト変換等	Text-to-Image生成、Image-to-Text生成、画像分類、物体検出、ビデオ生成、キャラクタ生成等	コード補完、対話的システム開発、コード解析、デバッグ、DevOps自動化、文書化等	仮想アシスタント、コミュニケーションロボット、顧客サービス、実行計画等	創薬、ゲノム解読、医療診断、個人教師等

様々なタスクで高い専門的能力を達成

- 対話AIのChatGPT、米名門大のMBA合格レベル
  https://www.nikkei.com/article/DGXZQOUC243WM0U3A120C2000000/
- ChatGPTが米医師資格試験で合格ライン
  https://medical.jiji.com/news/56105
- OpenAI「GPT-4」発表、司法試験で上位10%
  https://www.nikkei.com/article/DGXZQOGN1507H0V10C23A3000000/
- AI作品が絵画コンテストで優勝、アーティストから不満噴出
  https://www.cnn.co.jp/tech/35192929.html

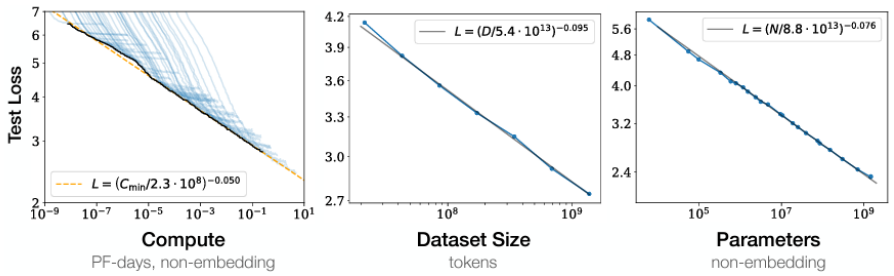
技術開発状況

- 基盤モデルは最近数年で急速に超大規模化、最先端モデルは学習1回の実行費用が数十億円
- 最先端の基盤モデル・生成AIの技術開発は米国ビッグテック企業が大きく先行
- 2023年はオープンソース化や応用開発、(中規模な)国産基盤モデルの後追い開発が活発化
- 産業・ビジネス用途だけでなく研究用の基盤モデル開発も立ち上がった(LLM-jp)

スケーリング則

https://doi.org/10.48550/arXiv.2001.08361

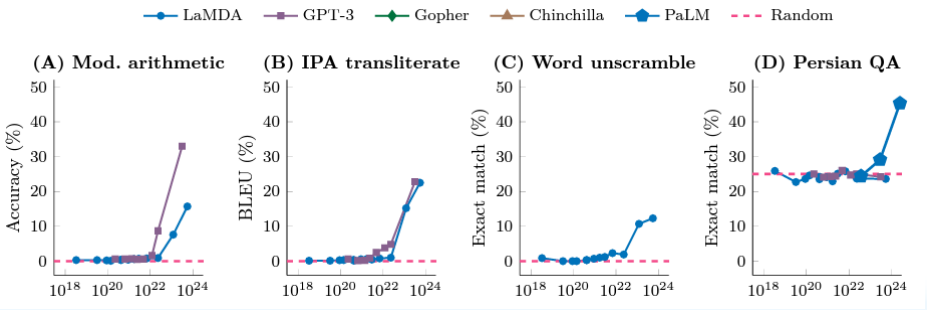
モデルが巨大化するほど性能が向上！



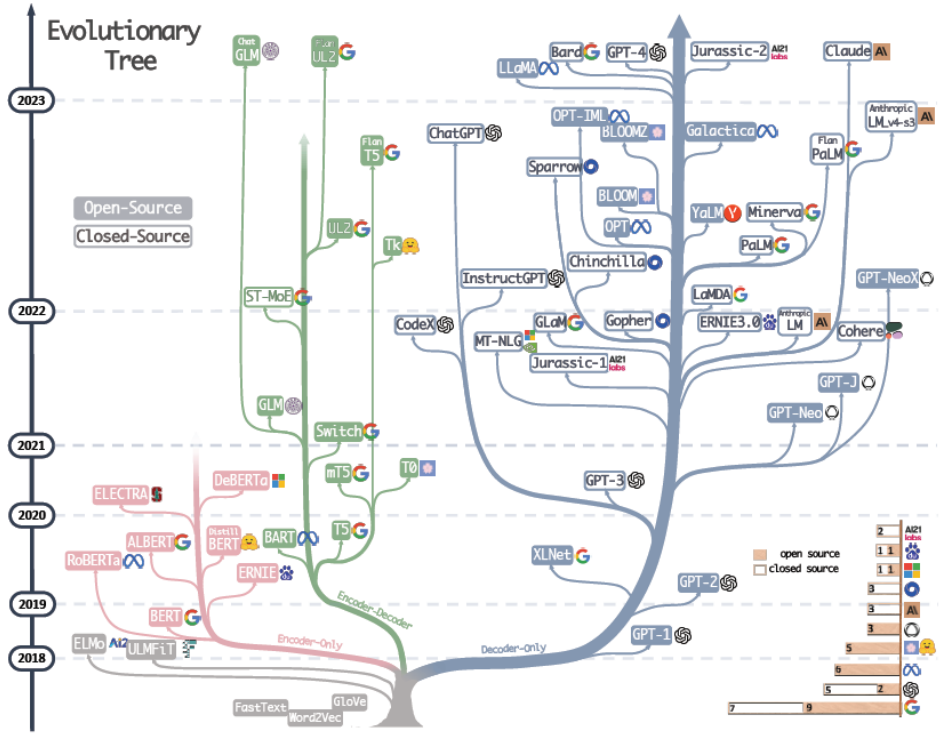
創発的現象

https://doi.org/10.48550/arXiv.2206.07682

モデルの規模がある点を超えると性能が劇的に向上



基盤モデル(LLM)の系譜



国産基盤モデル開発

[2020～2022年]
LINE+NAVER, rinna, ABEJA

[2023年～]
CyberAgent、オルツ、ZResearch、NEC、NTT、Preferred Networks、Turing、ソニー、ソフトバンクなど

研究開発用: LLM-jp (NIIを中心にオールジャパン体制), NICT, 東大松尾研, 東工大岡崎研・横田研＋産総研　ほか

中国での基盤モデル開発

BAAI(北京智源人工智能研究院)、Baidu、Alibaba等が開発：モデル規模では米国を上回るものも開発されているが、性能に関しては特段注目されるものは生まれていない

AIリスクの拡大

- ハルシネーション、社会的バイアス、情報漏洩、著作権・肖像権など、生成AIの出力に関わる様々な問題が顕在化
- フェイク生成や犯罪利用などの悪用問題、粗悪な生成AI乱立に対する第三者認証の必要性
- 海外ビッグテック企業サービスへの過度な依存は経済安全保障や産業競争力も左右

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**

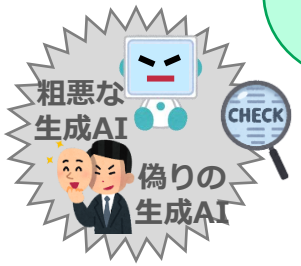
- **クリエーターや俳優・声優などの反発**、創作市場・文化への影響

- 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



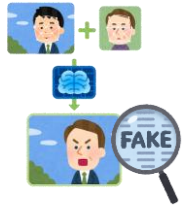
- 個人情報・機密情報の除外や**品質確保**をしていない
- **偽ってもバレない**だろう

- 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**

政策動向

■ イノベーション促進とAIリスク対処の両面から各国での政策検討と国際的議論
■ 特に2023年は生成AI対応の動きが国際的に活発化

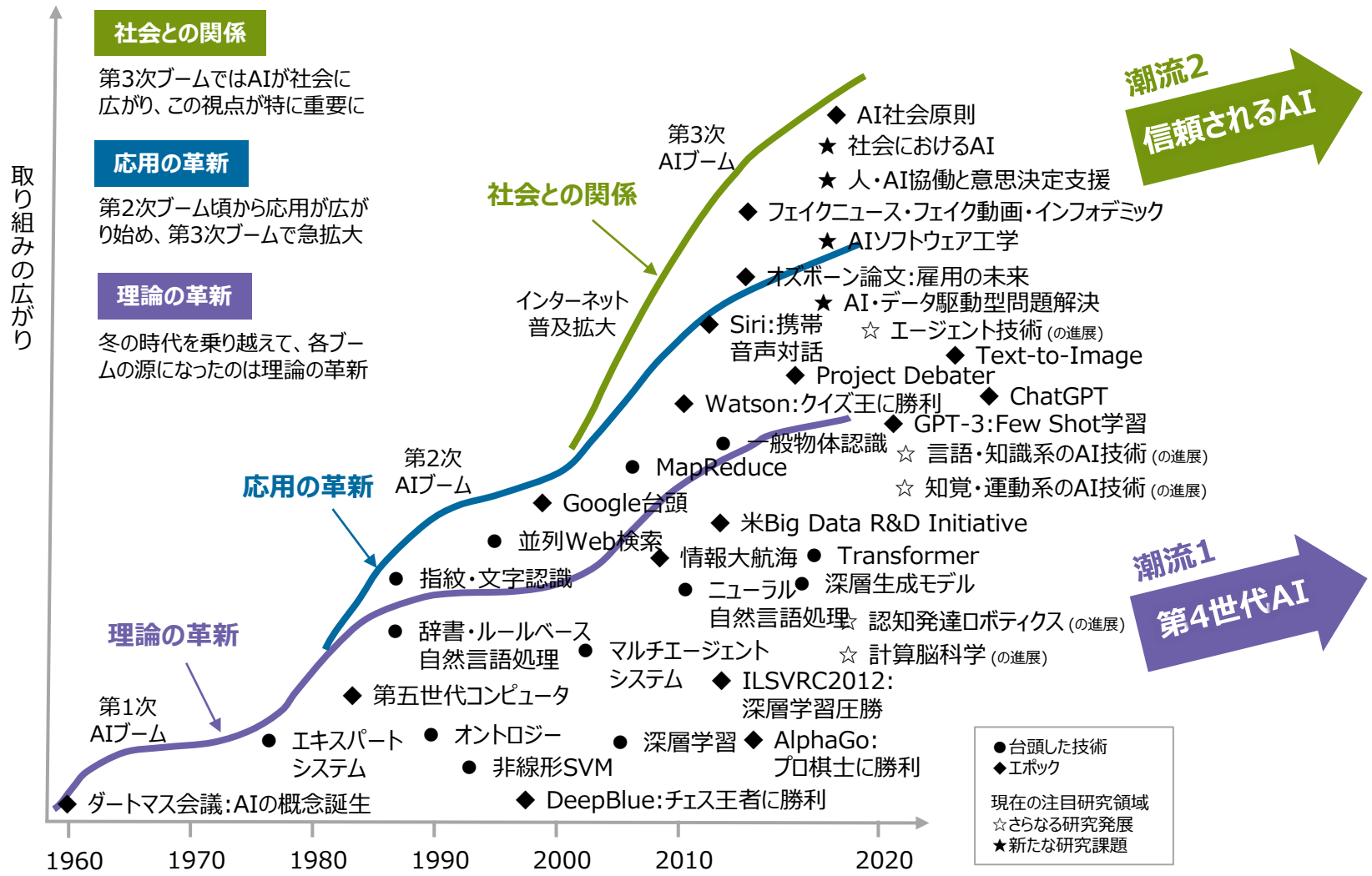
	政策面の特徴、技術強化・規制の傾向など	2023年の動き
米国	ビッグテック企業がビジネスと基礎研究の両面で圧倒的優位で、基盤モデル・生成AIの研究開発においても世界をリードしている状況であり、AI技術のもたらすベネフィットとリスクのバランスを法制度で調整しつつも、過度の規制によってイノベーションを阻害することは避けている。 ビッグテック企業やスタートアップによる民間の活発な技術開発の一方、DARPAなどの国の機関が中長期的な戦略投資を行い、経済・国家安全保障のためのAI強化も推進。	2023年1月 NIST AI Risk Management Framework 2023年5月 責任あるAIイノベーションの計画発表 2023年7月 責任あるAIの自主的コミットを7社と合意 2023年10月 AI安全に係る大統領令 2023年11月 米国AI安全研究所の設立表明
中国	国際学会でも躍進著しく、米中二強という状況だが、AIリード企業5社を選定するなど、政府がAI産業を後押しし、AI実装スピードに勢いがある。 政府は監視・管理社会の構築のためにAIを活用しており、他国と異なるAI応用技術開発や、生成AI規制も進めている。	2023年1月 ディープフェイク規制 2023年4月 生成AI規制 2023年10月 グローバルAIガバナンスイニシアチブ
欧州	各国のAI戦略に加えて、研究・イノベーションの枠組みプログラムによる国横断のAI研究（AI for Europe）を推進。 各個人の権利を重視し、法制度でAIをコントロールしようというハードロー指向で、人権や正義に根差した理念主導で国際的議論を進める傾向。AIに関わる国際ルール作りを通して米中・ビッグテック企業に対抗、EUルールをグローバル企業が遵守することでデファクト化につながる「ブリュッセル効果」を発揮。	2023年6月 AI法案修正(生成AI対応を盛り込み) 2023年7月 AI条約の統合版ドラフト公開 2023年11月 英国でAI安全サミットを開催(ブレッチリー宣言)、英国AI安全研究所を設立
日本	「人間中心のAI社会原則」を策定し、G20やOECDでのAI原則策定へ打ち込み。「AI戦略2019」で研究開発目標としてTrusted Quality AIを打ち出し、理研AIP・産総研AIRC・NICTを中核国研として国のAI研究を牽引する体制を整備。 リスクへの対応、AIの利用促進、AI開発力の強化を柱として、生成AIに関わる取り組みも進めている。 AIガバナンスでは、アジャイルガバナンス、ソフトローを指向。GPAI議長国であり、さらにG7議長国として「G7広島AIプロセス」を主導。	2023年4月 G7デジタル・技術大臣会合閣僚宣言 2023年5月 AI戦略会議発足 2023年10月 G7首脳共同声明・国際指針・国際行動規範 2023年12月 広島AIプロセス包括的政策枠組み 2023年12月 AI安全性評価機関をIPAに1月設立を表明

目次

時系列俯瞰図　AI研究の2つの潮流

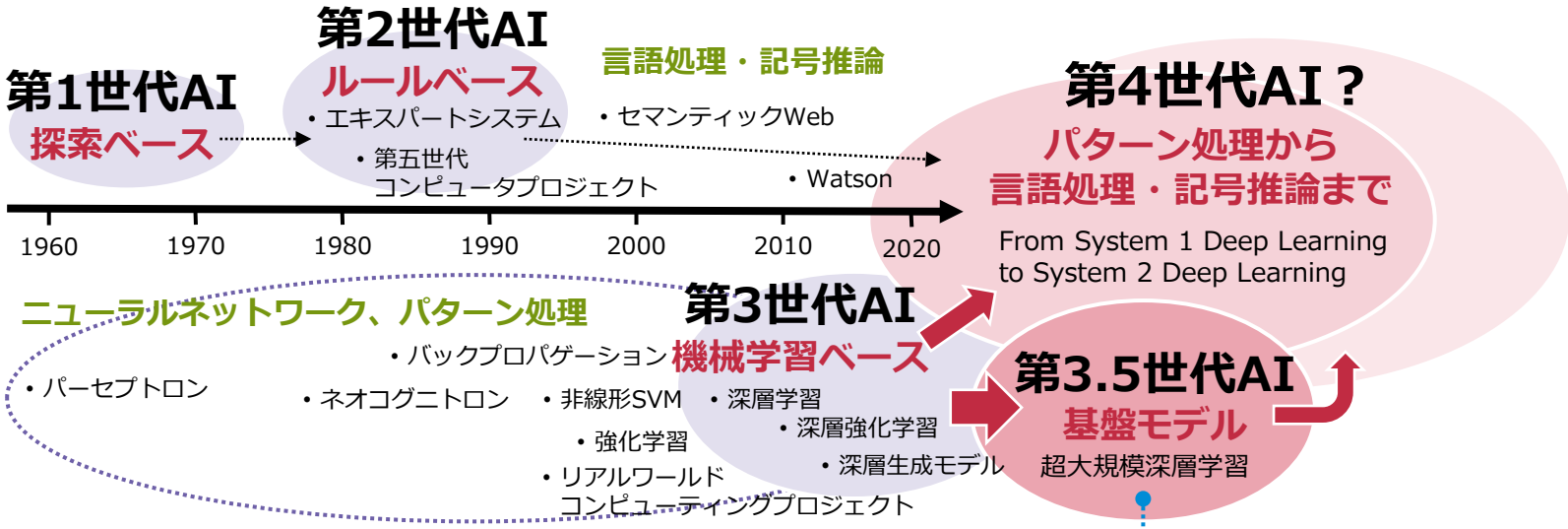
■ AIブーム3回の様相の違い
- **第1世代AI：**概念誕生、トイシステムレベル
- **第2世代AI：**ルールベース、形式知の記述、応用の始まり
- **第3世代AI：**深層学習ベース、暗黙知の学習、応用の拡大、社会への影響大

■ 現在の2つの潮流
- **第4世代AI：**
  AIの基本原理の進化
- **信頼されるAI：**
  社会からの要請の充足



第4世代AIへの技術発展

- 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・確率モデルであることによる限界・問題点が存在する
- これらの限界・問題点を克服する第4世代AIに向けたヒント・技術シーズ： 基盤モデルのメカニズム解明、人間知能に関する研究成果・知見、AIと人間・社会との関係性モデルなどの取り組み



現在の基盤モデルの問題点

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：**大きなタスクのサブタスク分解や論理構築・論理演算が苦手
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

第4世代AIに向けた技術シーズ・研究開発動向

(a)基盤モデルの仕組みをベースに外付け改良を積み上げる

プラグイン　苦手な処理を外部モジュールとして実装・連携(Walfram数式処理プラグインなど)

autoGPT　検索・ファイルI/O等も含むワークフローを目的に合わせて自動設計

LangChain　外部リソースと連携させるための仕組み

RAG (検索拡張生成)　プロンプトに応じて外部知識参照

ほか

(b)基盤モデルのメカニズムの解明に基づく基本原理の改良

LLM勉強会　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、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

(d)知能を人間・社会との関係性の側面から発展させる

コモングラウンドは、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背景だが、人と現在のAIの間ではそれが欠けている

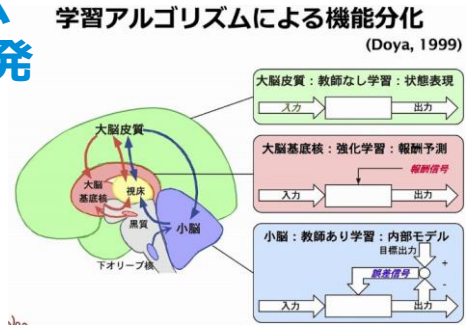
環境の中の知能　社会におけるAIの役割を、メタAI、キャラクターAI、スパーシャルAIという3種類で構成して人間とインタラクションするMCS-AI動的連携モデルなど

HAI　Human-Agent Interactionのモデルおよび設計論など

(c)人間の知能のメカニズムからヒントを得た新原理開発

脳科学(脳情報処理)や発達科学(認知発達ロボティクス)の発展による知能への構成論的アプローチ

知能に関する二重過程理論、幼少期の発達に関する予測符号化理論のAI応用など

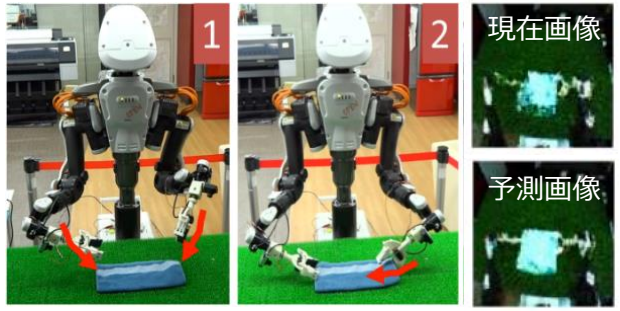


感覚信号と予測との誤差を最小化するように認知機能が発達

成長

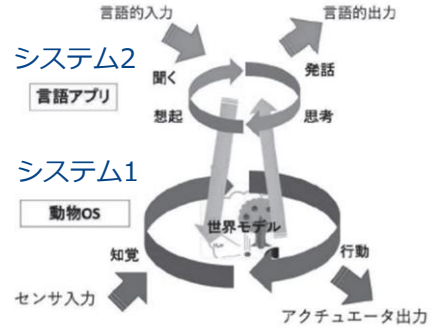
即応的システム1 ＋ 熟考的システム2

予測符号化理論



深層予測学習によるロボット制御
https://ieeexplore.ieee.org/document/7762066

二重過程理論



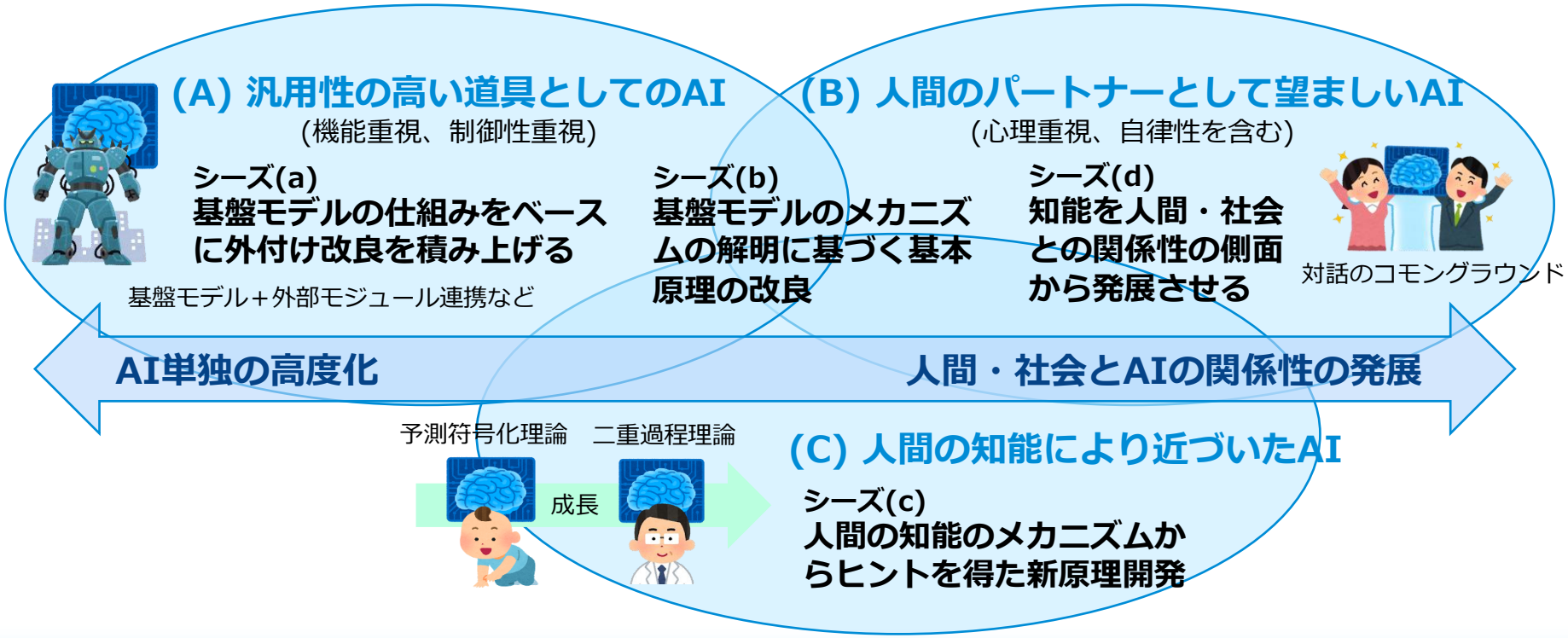
知能の2階建てモデル
[認知科学29(1), 2022]

第4世代AIに向けた研究開発の方向性

- 現状の各技術シーズは、問題点に対する克服の見込みが限定的あるいは未知
- 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**、という方向があるが、これらは徐々に融合する見通し
- アプローチ(技術シーズ)の可能性を幅広に認めつつ、融合・シナジーを生み出すことで、基盤モデルからさらに発展した次世代AIモデルの創出を加速

克服すべき問題点	技術シーズ(アプローチ)			
	(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（○/△）



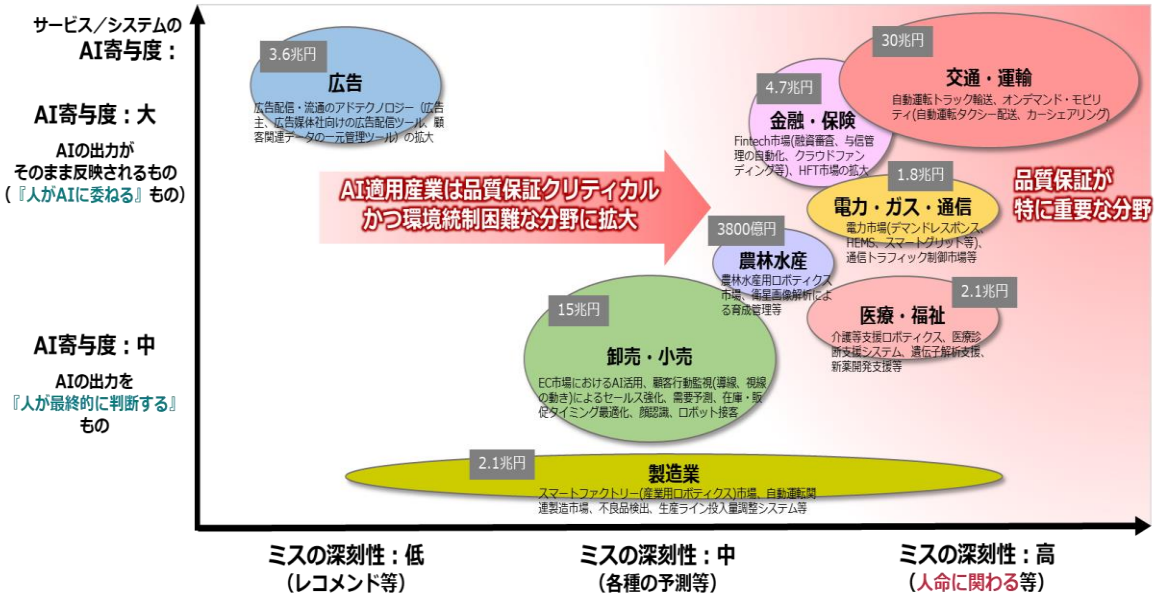
目次

信頼されるAIへの社会的要請

ブラックボックス問題、バイアス問題、脆弱性問題、品質保証問題等

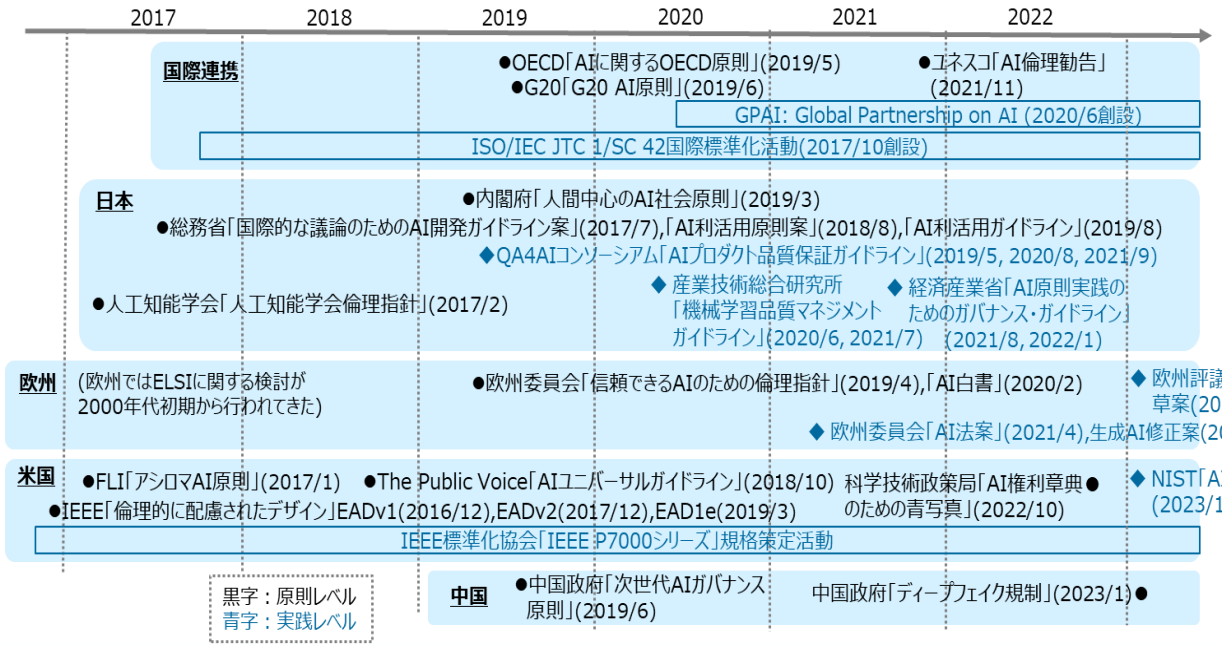
- 第3次AIブーム以降、AI応用が品質クリティカルな分野に拡大し、安全性・信頼性が重要課題に
- AIの開発・運用における社会原則・倫理指針は、2019年前後にまず国・国際レベルの議論を通して策定され、その後、原則から実践フェーズへ移行
- 2023年は生成AIがもたらすリスクに対する国際的な議論が活発化(G7広島AIプロセス、AI安全サミット等)、重点はAI倫理から脅威対策(予期せぬ事態への危機感、偽情報対策等)とAI安全への国際協力へ

AI応用分野の拡大



注：市場規模の金額は2030年のAI適用産業の予想市場規模　[出典] EY総合研究所：人工知能が経営にもたらす「創造」と「破壊」

AI社会原則・倫理指針の策定と実践

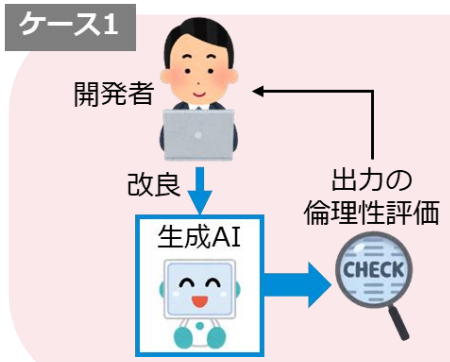


信頼されるAIの技術開発状況

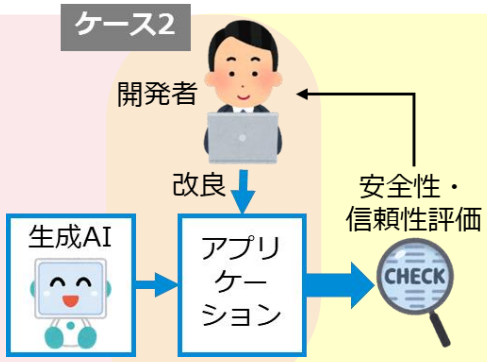
- **AI原則・指針を実現する技術開発：** 従来の演繹的開発法(プログラミング)に対して、機械学習応用は帰納的開発法であり、2018年頃から新たに機械学習工学(AIソフトウェア工学)が立ち上がった
  - 機械学習の品質管理・テスト、公平性評価、プライバシー保護、説明可能AI、脆弱性対策など
- **悪用面**では、2017年前後からディープフェイクが社会問題化、新種のサイバー攻撃手段化
  - メディア詳細解析によるフェイクメディア検出、電子透かし・ブロックチェーン等による出所・経路追跡等
- **生成AIの高品質化・利用拡大によって、問題が一層深刻化、かつ、リスクケースが拡大・多様化**

開発時

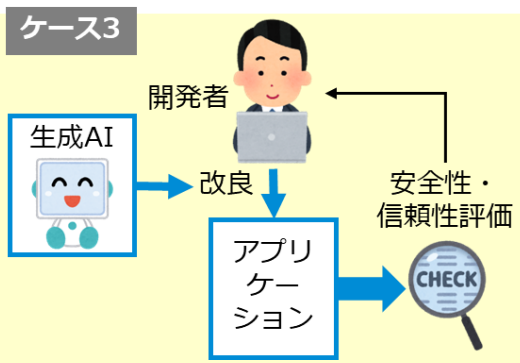
生成AI出力の倫理性確保(不適切な応答回避)



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

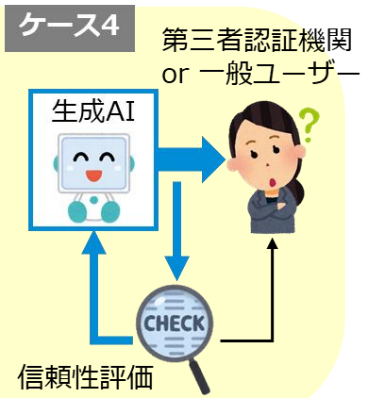


- 機械学習工学(AIソフトウェア工学)が新しい研究分野として立ち上がり、方法論・技術群が徐々に蓄積されつつあるが(QA4AI, AIQM, MLSEでの活動等)、さらに継続して強化必要
- 生成AI対応は、RLHF技術だけでなく、ソフトウェア工学面からのアプローチ強化も必要

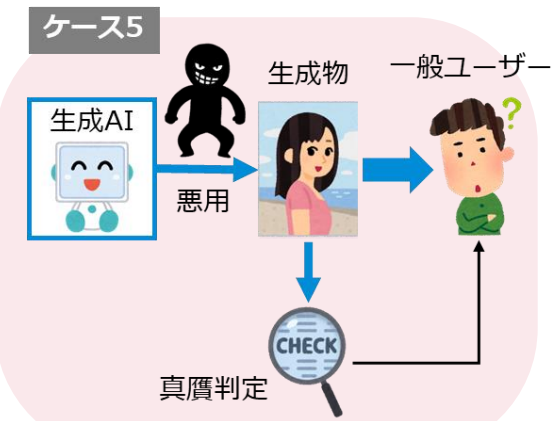
- 生成AI利用拡大によってリスクケース拡大
- AI安全性評価機関の設立等、問題は認識され、技術的な対策がこれからの重要課題

利用時

乱立した生成AIが良質なものかの審査など



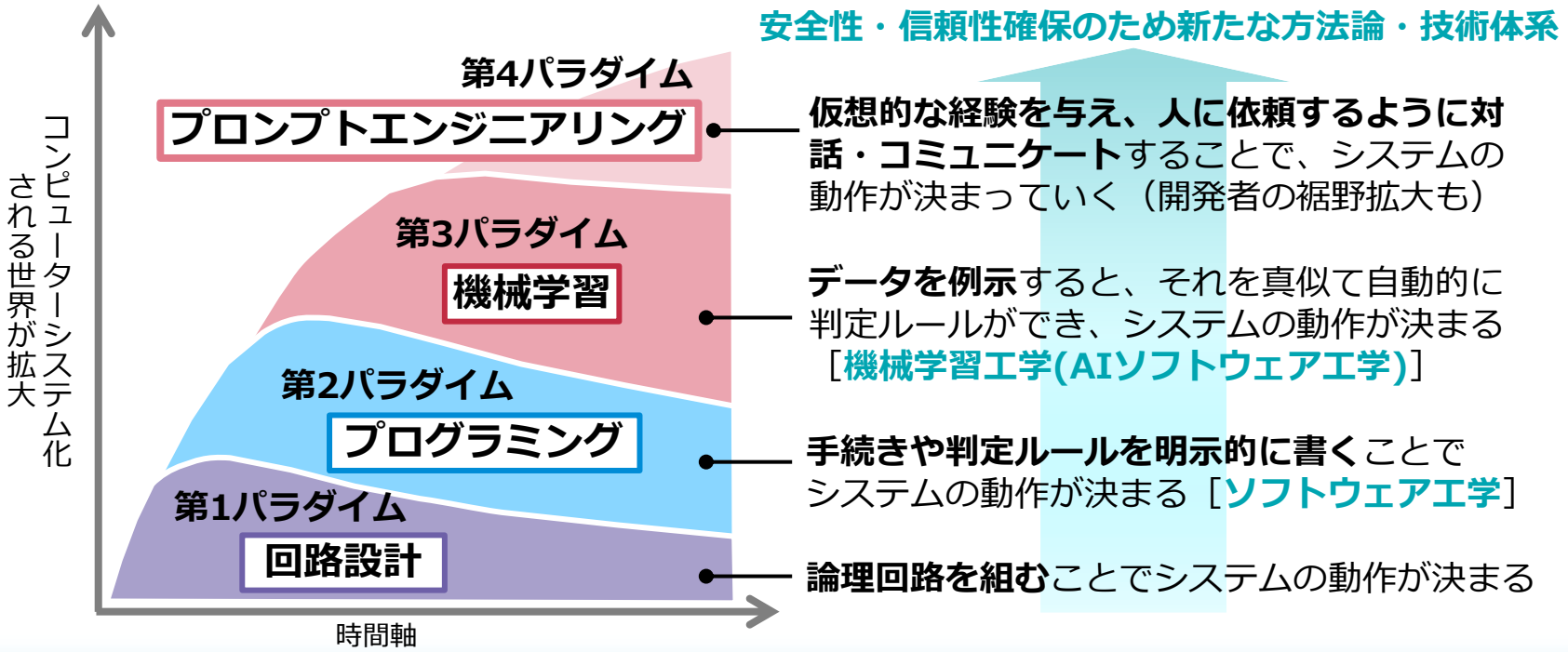
フェイクにだまされないようにする対策など



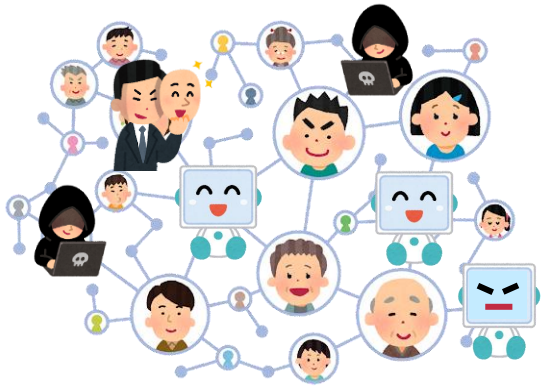
- フェイク問題が深刻化し、対策技術への取り組みが急速に活発化(CRESTFakeMedia, 米国DARPA MediaFor・SemaFor等)

信頼されるAIに向けた研究開発の方向性

- 各リスクケースについて対策技術の継続的な強化・整備は不可欠
  - 生成AI・プロンプトによるシステム開発は新たなパラダイムとしてソフトウェア工学的な方法論整備
  - 生成AIの悪用や増大する脅威への対策として、AIセキュリティ技術開発の強化・体系化
- さらに、個別対策の積み上げだけでなく包括的視点から、不安定なAIと不完全な人間が混在したマルチエージェント社会におけるAIリスク低減の新アプローチ創出も
  - マルチエージェント社会のメカニズムデザイン理論、社会的トラスト形成など



システム開発のパラダイムシフト



社会的よりどころの強化と
多面的・複合的検証

振る舞い予想・対応可能性
対象真正性
内容真実性
トラストの3側面

マルチエージェント社会のトラスト形成

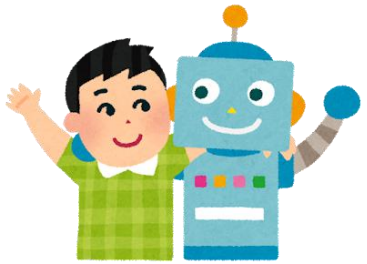
CRDS

第4世代AI＋信頼されるAI

■ 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、安全性の問題や、深刻化する様々なリスクが、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



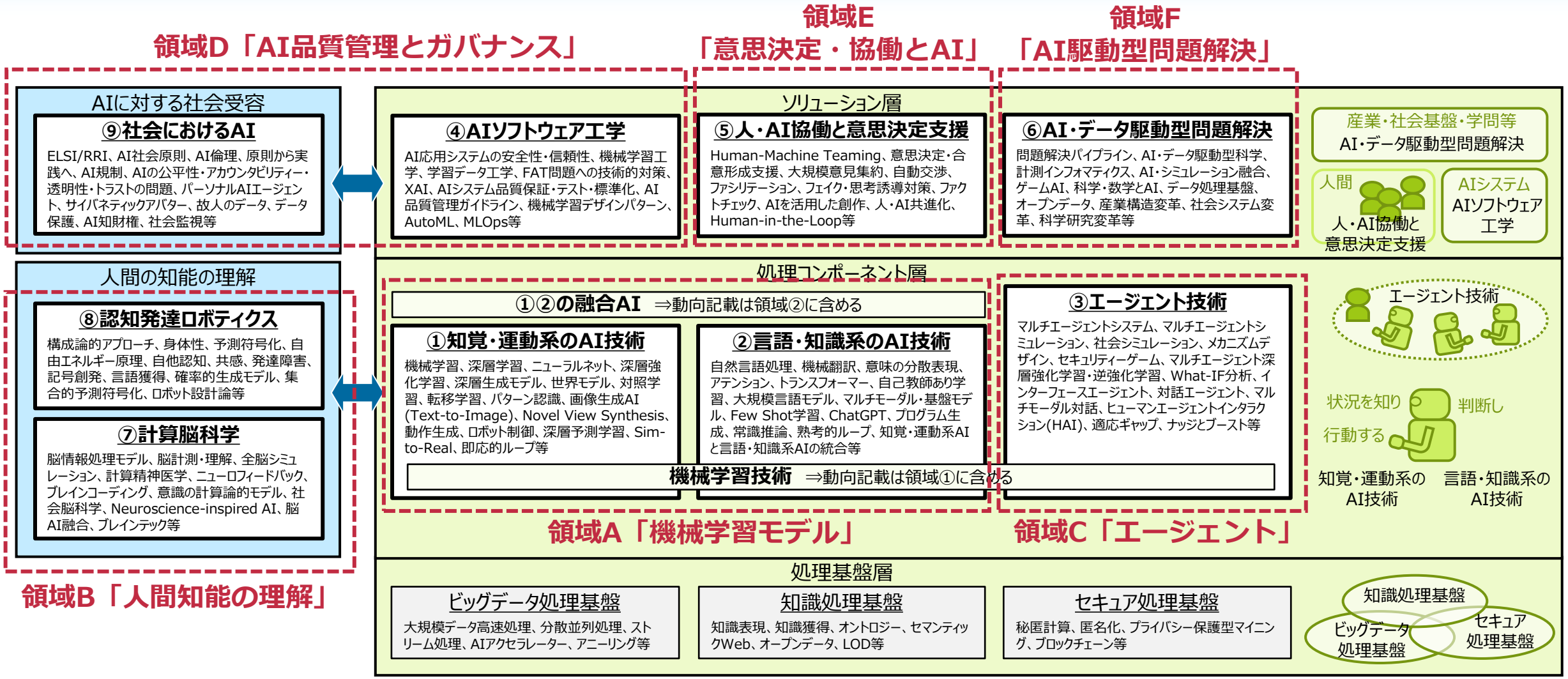
ハルシネーションの抑制やフェイクの判別が現在より進めやすくなり、不正確な情報や偽情報の流通による社会混乱や犯罪(詐欺・なりすましなど)の防止に役立つ

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

CRDS

目次

構造俯瞰図　　9つの研究開発領域 → 6領域にまとめ直して次頁以降で動向概説



領域A・B・Cが「第4世代AI」、領域D・E・Fが「信頼されるAI」の潮流に概ね対応

領域A「機械学習モデル」

研究開発の動向・方向性

- 2012年画像認識競技会ILSVRCで深層学習が大きな精度向上を示して以来、深層学習がAIの中核技術となった
- 言語系では当初精度が上がらなかったが、意味の分散表現、トランスフォーマー&アテンション機構、穴埋め型の自己教師あり学習、スケーリング則に基づく超大規模化により、驚異的な精度向上を達成
- 画像系では当初CNN型の深層学習が主流だったが、画像系でもトランスフォーマーが高い性能を示して移行
- 当初は識別モデルの深層学習が中心だったが、深層生成モデルの性能が急速に向上、画像系から言語系までトランスフォーマー型に統合され、汎用性・マルチモーダル性の高い基盤モデルが最先端モデルとして発展
- 画像等の知覚系から世界モデルを作り、言語系や動作系につなげる枠組みも検討が進んでいる

注目トピック

- 基盤モデル(超大規模深層学習)を用いた生成AIが高品質化、爆発的に普及
  - 2022年11月末に公開されたChatGPTのアクティブ利用者数が2ヶ月で1億人超え、自然で専門家並みの応答、様々な知的作業を変革
  - 画像生成はGANやVAEから拡散モデルに移行し、極めて高品質な画像生成がテキストから可能になり、Midjourney、DALL-E等、広く一般利用が広がる
  - 入力文(プロンプト)から文章・画像だけでなく、プログラムコードの生成等、多様な用途に利用拡大
- 深層学習をロボット制御に使う取り組みも注目されている
  - 予測符号化理論に基づく深層予測学習による、模倣からの柔軟な動作学習
  - 基盤モデルとロボットを組み合わせ、言語による曖昧な指示で動作(PaLM-SanCan)
  - ロボット実機での大規模なトランスフォーマー学習(RT-1/2/X)による動作生成汎化
- NeRF (Neural Radiance Field)
  - 複数視点画像から高品質な3D生成が可能

RT-1 https://robotics-transformer1.github.io/

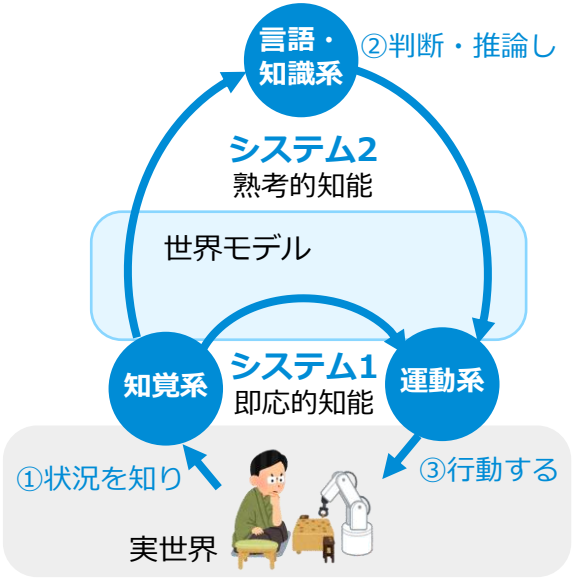
領域B「人間知能の理解」

研究開発の動向・方向性

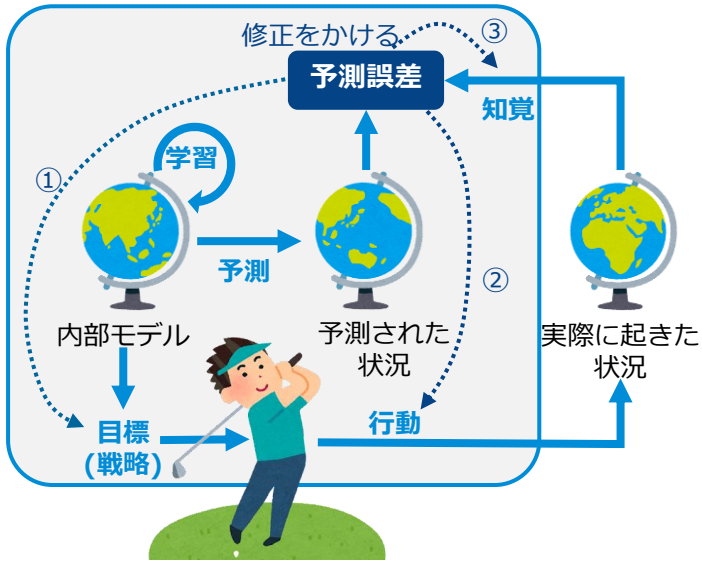
- **計算脳科学:** 脳情報処理の計測・理解技術が大きく発展(fMRI関連技術、ニューロフィードバック等)、脳情報処理からヒントを得たAIモデルの実績拡大(深層学習、強化学習、アテンション等)
- **認知発達ロボティクス:** 身体性・社会的相互作用を通した自律的な認知機能の発達過程を構成論的に理解する取り組み、予測符号化理論、記号創発、自他認知等のモデル化と、ロボットでの実装・検証が進む
- 発達障害、精神・神経疾患の解明や治療・予防への貢献も進められているほか、BMI等の計測・介入技術の様々な応用がニューロテックやブレインテックと呼ばれて活発化しつつある

注目トピック

- 次世代AIへのヒントとして特に注目される2つの理論、AI・ロボットへの実装も取り組まれている
  - 二重過程理論(即応的システム1＋熟考的システム2)
  - 予測符号化理論(予測誤差最小化原理、自由エネルギー原理)
  - 予測符号化の拡張として集合的予測符号化・記号創発システム
- 意識に関する計算論的モデル化も検討されている
  - 統合情報理論、グローバルニューロナルワークスペース理論等
- 富岳全脳シミュレーションプロジェクト
  - 世界で初めてニューロン・シナプスのヒト規模のシミュレーションを達成



二重過程理論



予測符号化理論
(予測誤差最小化原理、自由エネルギー原理)

領域C「エージェント」

研究開発の動向・方向性

- **マルチエージェントシステム:** 交通・電力・金融・災害対策・感染症等のシミュレーションで社会実装・活用が進展、深層学習を用いた自動メカニズムデザインやマルチエージェント逆強化学習等による高度化
- **インタフェースエージェント:** 対話エージェントはタスク指向型対話から非タスク指向対話(雑談)へ、大規模言語モデル(LLM)の適用が進む一方、言語・音声のみでなくマルチモーダル対話や、認知・心理面にも配慮したHuman-Agent Interaction (HAI)の設計法等の研究開発にも広がり

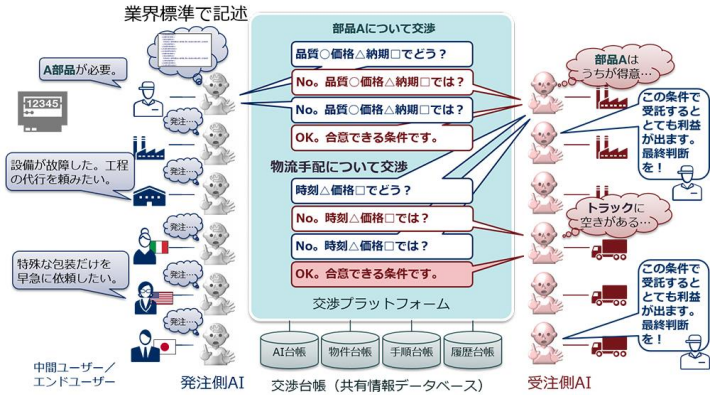
注目トピック

セキュリティゲームの実用成果拡大

- テロリストなどの攻撃者から空港などの重要施設を守るために、適切な警備員配置を決定する警備計画問題を、ゲーム理論で定式化(Stackelbergゲーム)
- 空港警備等の治安問題(ロサンゼルス空港に適用)、保全問題(世界100か所以上の国立公園に適用)をはじめ、公衆衛生問題、セイバーセキュリティ、森林保護、交通システム等への適用も

エージェント間自動交渉

- 予め定めた効用関数に基づいて、AIが高速に交渉を実行
- ドローン運行管理、SCM等で実証実験



https://jpn.nec.com/press/201908/20190821_02.html

Generative Agents実験

- 生成AI(GPT-4)で作った25体のAIエージェントを、仮想空間内で対話しながら生活させた実験

https://doi.org/10.48550/arXiv.2304.03442

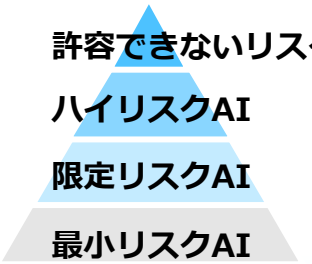
領域D「AI品質管理とガバナンス」

研究開発の動向・方向性

- 機械学習によるシステム開発のパラダイム転換(演繹型→帰納型)に対して、2018年頃から新たな方法論・技術群の研究開発が立ち上がった(機械学習工学/AIソフトウェア工学/Software 2.0等と呼ばれる)
- AI品質・安全性等を扱う国際標準化活動(ISO/IEC JTC 1/SC 42等)、自動運転分野のAI安全規格検討が進展
- 2019年前後にAI倫理・社会原則が国・国際レベルで議論・策定、その後、原則から実践へフェーズが移行、さらに生成AIの台頭を受けて2023年には脅威への対策や安全性評価が国際的論点に
- 社会におけるAIの課題抽出・目標設定、その実現のための制度設計と技術開発、それらを実践・評価して全体統治するAIガバナンスの在り方が重要課題になるとともに、社会的トラスト形成の観点からの検討も

注目トピック

- 欧州AI法案がAI規制を先導
  - 2019年4月に欧州委員会が公表
  - AIをリスクに応じて4レベルに分けて規制の仕方を設定
  - 2023年6月に生成AI対応も盛り込み修正



→利用禁止
→事前適合性評価
→透明性義務
→規制なし

- AIリスクに対して2023年広島AIプロセスとAI安全サミット
  - 日本はG7議長国として広島AIプロセスを主導、10月G7首脳共同声明・国際指針・国際行動規範、12月に広島AIプロセス包括的政策枠組みを公表
  - 英国主導で11月にAI安全サミット開催、ブレッチリー宣言として、予期せぬ事態への危機感、偽情報対策、AI安全への国際協力を強調
  - AI安全性評価機関として、英国は11月にAI安全研究所を設立、米国も同様の機関設立表明、日本もIPAに1月設立を発表
- AI応用システムの品質管理ガイドライン
  - 日本ではソフトウェア工学面からの実践的取り組みが国際的には早い時期(2018年)から立ち上がった(機械学習工学研究会MLSE)
  - 実践レベルの詳細なガイドラインとして、QA4AIガイドライン(QA4AIコンソーシアム)、AIQMガイドライン(産総研)が知られており、国際標準化への撃ち込みも進められている

領域E「意思決定・協働とAI」

研究開発の動向・方向性

- 情報氾濫に伴う可能性の見落としやフェイク生成等による情報操作によって判断ミスを起こすリスクの高まりに対して、意思決定を支援する様々な取り組み
- 何らかの目的達成に向けた人とAIが協力して取り組むHuman-AI Teamingも重要課題に

意思決定支援技術の研究開発動向

研究開発課題	取り組み状況・方向性
膨大な可能性の探索・評価	マルチエージェントシミュレーション、因果推論・探索、常識推論等
自動意思決定・自動交渉	機械学習・最適化、自動交渉エージェント等
大規模意見集約・合意形成	ファシリテーションエージェント、メカニズムデザイン等
多様な価値観の把握・可視化	言論マップ生成、議論マイニング、VR・ゲーミングによる追体験等
フェイク対策	ソーシャルネット分析、フェイク検出、ファクトチェック支援等
意思決定に関する基礎科学	脳の意思決定メカニズム、行動経済学、ELSI・社会受容性等

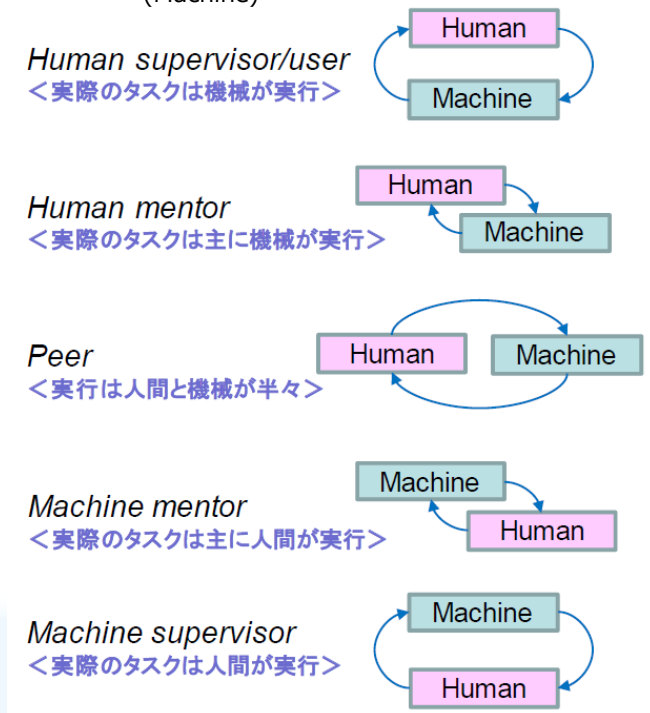
注目トピック

- 生成AI高度化によりフェイク対策が喫緊の課題となり、技術開発活発化
  - 米国DARPA: Media Forensics (MediFor)、Semantic Forensics (SemaFor)
  - JST CREST FakeMedia、NIIシンセティックメディア国際研究センター(SYNTHETIQ VISIONをライセンス事業化)

https://research.nii.ac.jp/~iechizen/synmediacenter/synthetiq/index.html

- AI活用創作や人・AI協働創作の探究が現場と研究開発の両面で進展
  - 2015年頃からGAN等を使った絵画生成・音楽生成が始まったが、2022年の生成AI公開で一気に創作現場での利用進展(一方で作風を模倣されたクリエータからの反発も)
  - 生成AIを用いたキャラクターデザイン、ストーリー生成、音楽生成等、制作効率化に活用
  - ゲーム開発現場での活用活発化
  - 2023年「TEZUKA2023」プロジェクト(NEDO事業の中で実施)により、人・AI共創によるマンガ「ブラックジャック」の新作が制作・公開された
    https://www.nedo.go.jp/news/other/ZZCD_100061.html
    (作品自体と共創プロセスの解説もWeb掲載されている)

Human-AI Teamingの5つのパターン

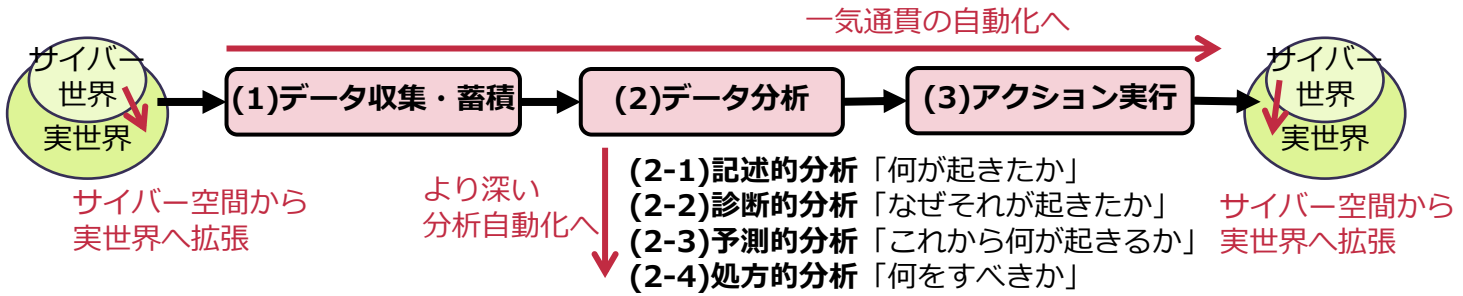


領域F「AI駆動型問題解決」

研究開発の動向・方向性

■ AI・ビッグデータ解析の技術発展によって大規模複雑タスクの自動実行や膨大な選択肢の網羅的検証が可能になり、問題解決手段の質的変化、産業構造・社会システム・科学研究などの変革へ

●問題解決パイプラインの技術発展:



●サイバーフィジカルシステムの技術発展:
- IoTデバイスの軽量化、省エネ化、高感度化、高解像度化、スマート化

●データ基盤の技術発展:
- データ処理基盤技術(大規模高速処理、分散並列処理、ストリームデータ処理等)
- データ保護技術(データ匿名化、分散プライバシー、秘密計算等)
- オープンデータ技術(LOD、データ連携基盤、共通語彙基盤等)

●計測の高次化の進展:
- 狭義の計測から広義の計測へ(物理量計測の高性能化、意味的計測、自律的計測、社会計測)

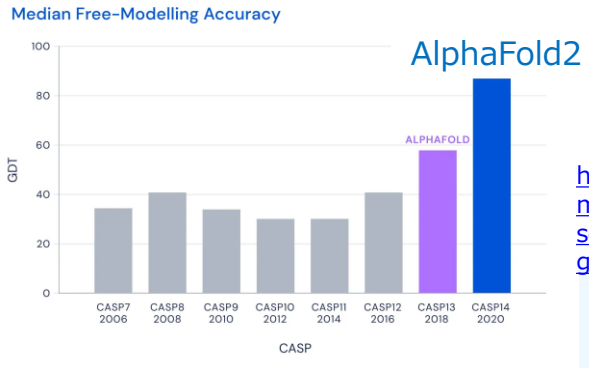
注目トピック

■ AI駆動科学プロジェクト発足相次ぐ
- Nobel Turing Grand Challenge
- 英国Alan Turing Institute
- ムーンショット事業、未来社会創造事業
- 生命科学分野、材料科学分野で先行

■ 深層学習の科学分野適用が大きく進展
- タンパク質構造予測問題(AlphaFold2)
- タンパク質言語モデル(トランスフォーマー適用)
- 数学の問題: 行列積計算アルゴリズム発見(AlphaTensor)、Ramanujan Machine、Minerva
- 学習物理学の創成(学術変革領域A)

■ AI・シミュレーション融合
- 汎用原子レベルシミュレーターMatlantis

■ ゲームAIの進化
- 完全情報ゲーム(囲碁等)から不完全情報ゲーム(ポーカー、リアルタイムストラテジーゲーム等)へ



https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

目次

基盤モデルの研究開発課題の全体観と戦略提言のターゲット

- **現状認識：**日本国内では基盤モデル・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おり、国際的には喫緊の問題への対策・ルール整備なども進められている状況
- **戦略提言のターゲット：**活発化している基盤モデル・生成AIの後追い開発や応用開発にとどまらず、その先の次世代AIモデルを創出する基礎研究の戦略強化を狙う

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

基礎研究振興部会(AI for Science)に取り込み ③
- 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

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

AIリスク対処研究 ②
- 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

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モデル研究 ①
- 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 ↕ 共通基盤

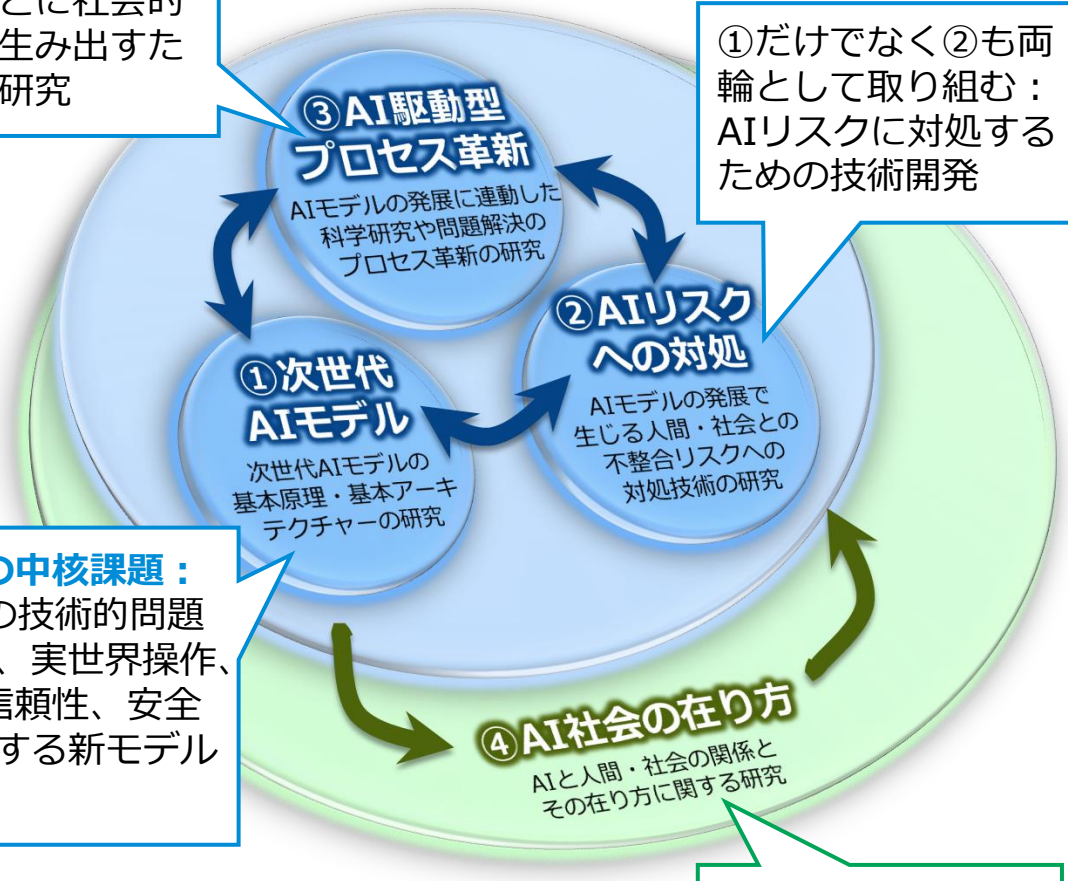
実務 ←→ 学術

基礎研究として重点的に取り組むべき課題
(戦略提言対象)

①は「第4世代AI」の潮流、
②は「信頼されるAI」の潮流に沿う
③はそれらの活用場面

重点的な研究開発課題(案)

①②をもとに社会的な価値を生み出すための基礎研究



①だけでなく②も両輪として取り組む：AIリスクに対処するための技術開発

基礎研究の中核課題：
現在のAIの技術的問題(資源効率、実世界操作、論理性、信頼性、安全性)を克服する新モデル研究

AI社会の在り方を人文・社会科学の側から考える

①次世代AIモデルの基本原理・基本アーキテクチャーの研究
- 望ましいAIの在り方を実現する新しい原理・アーキテクチャーの設計
- 現在のAI (基盤モデル、生成AI)のメカニズム解明と問題点に対する改良
- 人間の脳情報処理・認知発達過程の理解とそれに基づくAIの原理設計
- 身体性を含む実世界や他者との関係性に基づく知能モデル　など

②AIモデルの発展で生じる人間・社会との不整合リスクへの対処技術の研究
- 現在および次世代のAIモデル自体の倫理性・安全性の確保技術
- 生成AI・プロンプトを用いて開発される応用システムの安全性・信頼性の確保技術
- 利用側からのAIモデルの品質検証技術
- 多面的・複合的なフェイク対策技術
- 多数のAIと人間が混在・共生する社会で発生するリスクを低減する技術など

③AIモデルの発展に連動した科学研究や問題解決のプロセス革新の研究
- 現在および次世代のAIモデルを導入したAIロボット駆動科学による科学研究プロセスの高度化・自動化
- 問題の種類に応じたAIと人間の最適な協働形態の選択
- 大規模で複雑な問題の処理可能なサブ問題への分割　など

④AIと人間・社会の関係とその在り方に関する研究
- AI技術の発展が人間・社会にもたらす影響(正負両面)やリスク要因の推定
- リスク回避や社会受容のシナリオ作成、技術と制度の発展状況や社会・個人の価値観によって変化し得る中で望ましい関係の設計
- 望ましいAIの設計指針の導出　など

推進方策(案)

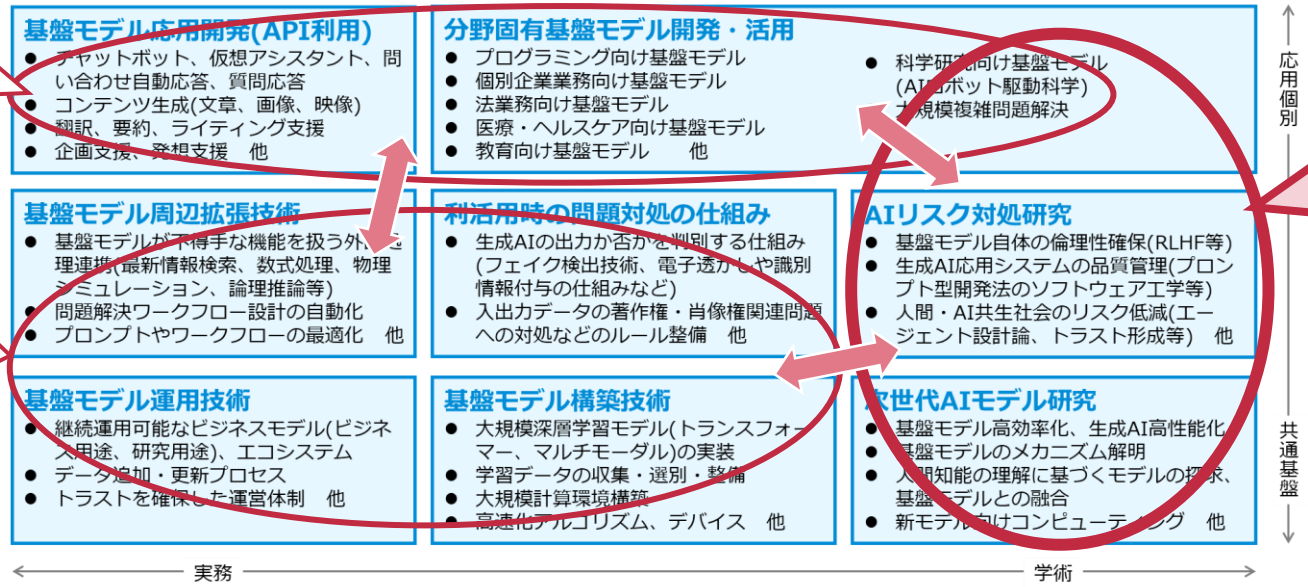
(1)基盤モデル・生成AI活用によって生産性向上DX・産業成長を促進

▶産業界主導(活用は政府・大学などでも活発に)

(3)後追いで追い抜くことは難しいが、(1)(2)を支え連動しつつ、徐々に底上げ

▶ビジネス用：産業界主導 (必要に応じて国が促進策)

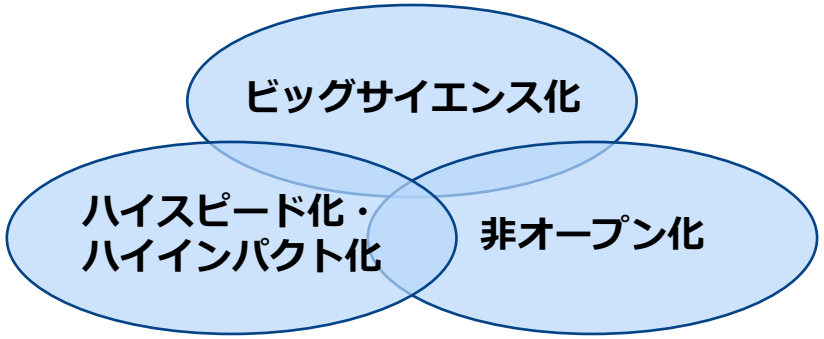
▶研究用：国の支援による共同利用施設



(2)次の世代のAIモデルで先行を狙う基礎研究を推進

▶アカデミア(大学+企業の基礎研究部門)が牽引、国のファンドで加速

研究開発形態の変化



研究開発形態の変化を踏まえた研究開発体制・基盤、プログラム、エコシステム

a. さまざまな研究機関が結集し協力し合う研究エコシステムの形成
b. 大規模計算機の共同利用施設の継続的な運用・強化
c. AIモデルとマルチモーダルデータの集約・共有・管理体制の整備
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
e. 情報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
f. 研究エコシステムのハブ機能を担う組織の設置
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
h. 研究エコシステムを支える人材の確保・育成