提言・提案

CRDS-FY2023-SP-03

戦略プロポーザル

次世代 AI モデルの研究開発

STRATEGIC PROPOSAL

Research and Development on Next Generation AI Models

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

研究開発戦略センター（CRDS）は、国の科学技術イノベーション政策に関する調査、分析、提案を中立的な立場に立って行う公的シンクタンクの一つで、文部科学省を主務省とする国立研究開発法人科学技術振興機構（JST）に属しています。

CRDSは、科学技術分野全体像の把握（俯瞰）、社会的期待の分析、国内外の動向調査や国際比較を踏まえて、さまざまな分野の専門家や政策立案者との対話を通じて、「戦略プロポーザル」を作成します。「戦略プロポーザル」は、今後国として重点的に取り組むべき研究開発の戦略や、科学技術イノベーション政策上の重要課題についての提案をまとめたものとして、政策立案者や関連研究者へ配布し、広く公表します。

公的な科学技術研究は、個々の研究領域の振興だけでなく、それらの統合によって社会的な期待に応えることが重要です。「戦略プロポーザル」が国の政策立案に活用され、科学技術イノベーションの実現や社会的な課題の解決に寄与することを期待しています。

さらに詳細は、下記ウェブサイトをご覧下さい

https://www.jst.go.jp/crds/

エグゼクティブサマリー

本プロポーザルでは、急速な技術発展を示し、大きな社会インパクトをもたらしつつある人工知能（AI）技術について、特に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提言を行う。

AI技術の発展は著しく、特に2022年11月末に登場したChatGPTをはじめとする、超大規模深層学習で作られた基盤モデルに基づく生成AIは、極めて自然な対話応答性能や高い汎用性・マルチモーダル性を示すようになった。これが人間の知的作業全般に急速な変革をもたらし、産業、研究開発、教育、創作などさまざまな分野に幅広く波及し、大きな社会インパクトが見込まれている。 McKinsey & Companyによると[1)]、生成AIは世界経済に年間数兆ドル相当の価値をもたらす可能性がある。わが国では特に、労働人口減少に対する生産性向上や産業・経済の活性化につながるという期待が大きく、米国ビッグテック企業が大きく先行する中、後追い開発や応用開発が活発化している。

一方、フェイクやなりすましなどへの悪用、ハルシネーション（もっともらしくウソを返す現象）、社会的バイアスの増長、著作権侵害をはじめ、ELSI（倫理的・法的・社会的問題）面から種々の懸念が指摘されている。また、さまざまな活動に用いられる汎用基盤技術となることから、海外のビッグテック企業提供サービスへの過度な依存は、経済安全保障面や科学研究・産業の国際競争力の面でリスクとなる。

以上のような期待と懸念に対して、イノベーションの推進策とリスクに対するルール作りの両面から、各国で生成AI関連の政策立案が急速に進んだ。 AIに関わる社会原則は、2019年に日本政府の「人間中心のAI社会原則」をはじめ国・国際レベルでの議論が進み、OECD原則（OECD Principles on AI）も策定された。原則から実践へとフェーズが移行し、現在は生成AIに関わるルール作りが重要な論点となっている。日本が議長国を務めたG7の広島AIプロセスでもこれが議論された。

このように、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、政策面ではその後押しや喫緊の問題への対策・ルール整備などが進められている。これに対して、本プロポーザルでは、さらにその先の次世代AIモデルを創出する基礎研究にフォーカスし、中長期を見据えた戦略強化を狙う。

この戦略立案に際して、関連動向の俯瞰的調査、50名を超える専門家へのインタビュー、ワークショップや学会などでの意見交換・議論を通して有望な方向性を見極め、以下を方針として特定した。

【方針A】次世代AIモデルへの発展を、AI単体の機能・性能の発展だけでなく、AIと他者（人間および他のAI）や社会との関係の発展にまで広げて考える。なぜならば、現在のAIモデルは、資源効率、実世界操作、論理性、信頼性、安全性などに問題を抱えている一方、驚異的な性能を示しているがゆえに人間の活動や社会の仕組みにも変化をもたらしつつある。現在の問題に対する改善だけでなく、AIと人間・社会の関係の在り方からも、次世代AIモデルの要件を考えていくことが必要になる。

【方針B】次世代AIモデルに向けた多様なアプローチを認めつつ、それらの間で問題意識や知見の共有を促す。技術ブレークスルーの可能性を幅広に探りながらも、ばらばらな取り組みとせず、シナジーや融合を促すことが、基礎研究を加速することになるからである。

【方針C】日本が抱える社会課題の解決や日本の特性に即した社会発展を支えるためのAI技術は自国で保有すべきである。他国のビッグテック企業と競争するのではなく、わが国として経済安全保障を確保しつつ、労働人口減少に直面する日本社会の発展や人々のWell-beingのために、現在のAIモデルの限界・問題点を克服することが必要である。

【方針D】ビッグサイエンス化などAI研究の形態が変化し、基礎研究においてもビッグテック企業が大きく

先行していることや、総合知による取り組みが不可欠になっていることを踏まえた、基礎研究推進の新しい体制づくりや研究エコシステムづくりを考える。

以上の方針に基づき、次世代AIのための基礎研究において取り組むべき重要な研究開発課題として、以下の①～④を挙げる。

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
② AI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
③ AI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
④ AIと人間・社会の関係とその在り方に関する研究

①が次世代AIモデル設計の中核になるが、②のAIリスクへの対処技術にも同時に取り組むことが不可欠である。①と②はいわば車の両輪のような関係となる基礎研究だが、そこから社会的な価値を生み出すための基礎研究が③のプロセス革新研究である。また、①②③は主に情報科学技術の研究であるのに対して、④は人・AI共生社会の在り方に関する人文・社会科学の研究である。④によって①②③の技術要件・指針が定まる一方、①②③の技術発展の結果として④の考え方や目標が変わり得る。これら四つは、相互に影響し合い、連動しながら進展するものである。

これらの取り組みを推進するにあたって、方針Dで触れたような研究開発形態の変化を踏まえる必要がある。すなわち、AI分野の研究開発は、基礎研究においても、ビッグサイエンス化、ハイスピード化・ハイインパクト化、非オープン化の傾向が強まっており、このような状況下での研究開発体制・基盤やプログラムの在り方を考え、それを支える研究エコシステムをつくることが必要である。その中には、以下のような要素が含まれる。

a. さまざまな研究機関・組織が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・評価・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. 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
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

Executive Summary

This proposal is related to the artificial intelligence (AI) technology, which is showing rapid technological development and having a large social impact, and provides strategic recommendations for basic research to create next-generation AI models that go beyond follow-on development and application development of the current foundation models and generative AI, which are becoming particularly active.

AI technology has developed remarkably, and in particular, generative AI based on the foundation model created by very large-scale deep learning, such as ChatGPT, which appeared at the end of November 2022, has come to exhibit extremely natural dialogue response performance and high versatility and multimodality. This is expected to have a significant social impact as it brings about a rapid transformation of human intellectual work in general and has a broad ripple effect in various fields, including industry, R&D, education, and creative work. According to McKinsey & Company[1)] , generative AI could bring trillions of dollars worth of value to the global economy annually. For Japan, where the working population is shrinking, it is greatly expected to increase productivity and bring about industrial and economic growth, and so follow-on development and application development is gaining momentum while U.S. big tech companies are far ahead.

On the other hand, various ELSI (Ethical, Legal, and Social Issues) concerns have been raised, including misuse for faking and spoofing, hallucination (the phenomenon of plausibly returning lies), increased social bias, and copyright infringement. In addition, since it will be a general-purpose fundamental technology used in various activities, excessive reliance on services provided by foreign big tech companies will create risks in terms of economic security and the international competitiveness of scientific research and industry.

In response to the above expectations and concerns, policy formulations related to generative AI were rapidly launched in various countries, both in terms of measures to promote innovation and rulemaking against risks. Social principles related to AI have been discussed and formulated at the national and international level in 2019, including the Japanese government's "Principles of Human-centric AI Society" and the OECD Principles on AI. After that, the phase is shifting from principles to practice, and now rulemaking related to generative AI has become an important issue. This was also discussed during the G7 Hiroshima AI process, which Japan chaired.

As mentioned above, efforts to follow-on development and application development of the current foundation models and generative AI are gaining momentum, and policy efforts to develop rules to address pressing issues are underway. In light of this situation, this proposal focuses on the strategic enhancement of basic research to create the next-generation AI models that go even further.

In developing this strategy, we set the following courses of action by identifying promising directions through a bird's-eye view of related trends, interviews with more than 50 experts, and exchanges of opinions and discussions at workshops and academic conferences.

A. We consider the development toward the next-generation AI models not only in terms of

the development of the functionality and performance of the AI itself, but also in terms of the development of the relationship between the AI and others (humans and other AIs) or society. Because, while current AI models have problems with resource efficiency, real-world operation, logic, reliability, and safety, they show phenomenal performance and are changing human activities and social structures. It will be necessary to consider the requirements for the next-generation AI models not only in terms of improvements to current problems, but also in terms of the nature of the relationship between AI and humans or society.

B. We acknowledge the diversity of approaches toward the next generation AI models while encouraging the sharing of awareness and knowledge of the issues among them. This is because exploring a wide range of possibilities for technological breakthroughs while encouraging synergy and fusion, rather than disparate efforts, will accelerate basic research.

C. AI technologies to solve social issues facing Japan and to support social development taking advantage of Japanese characteristics should be owned by Japan. Rather than competing with big tech companies from other countries, it is necessary for Japan to overcome the limitations and problems of the current AI model for the development of Japanese society and people's wellbeing in the face of a declining workforce, while ensuring economic security for our country.

D. We try to design a new research ecosystem to promote basic research, considering that the style of AI research is changing, such as the shift to big science, and that big tech companies are taking a large lead in basic research, and that efforts based on comprehensive knowledge are becoming indispensable.

Based on the above courses of action, the following (1) to (4) are important R&D issues that should be addressed in basic research for next-generation AI.

（1） Research on basic principles and basic architecture of next-generation AI models.

（2） Research on technologies to deal with the risk of incompatibility with humans and society arising from the development of AI models.

（3） Research on scientific research and problem-solving process innovation linked to the development of AI models.

（4） Research on the relationship between AI, humans, and society and how it should be.

While (1) is the core of the next-generation AI model design, it is essential to simultaneously address (2), techniques for dealing with AI risks. While (1) and (2) are basic research that are inseparable, (3) is process innovation research, which is basic research to create social value from (1) and (2). In addition, (1), (2), and (3) are mainly research on information science and technology, while (4) is research in the humanities and social sciences on the state of AI society. While the technical requirements and guidelines for (1), (2) and (3) are established by (4), the concept and goals of (4) can change as a result of the technological development of (1), (2) and (3). These four factors are interrelated and progress in tandem.

In promoting these efforts, it is necessary to take into account the changes in R&D style as mentioned in Course of action D. In other words, R&D in the field of AI, even in basic research, is becoming increasingly big science, high-speed/high-impact, and closed. It is necessary to consider how the R&D system, infrastructure, and programs should be under these circumstances, and to

create a research ecosystem that supports them. These include the following elements.

a. Formation of a research ecosystem where various research institutions cooperate with each other.
b. Continued operation and enhancement of large-scale computing facilities for shared use.
c. Developing a system for aggregating, sharing, evaluating and managing AI models and multimodal data.
d. Open international collaboration in basic research and rulemaking and the establishment of a support system.
e. Program design to promote the active participation of not only researchers in the technical fields but also in the humanities and social sciences.
f. Establishing an organization to serve as a hub for the research ecosystem.
g. Flexible and agile program management that leverages the research ecosystem.
h. To secure and develop human resources to support the research ecosystem.

目次

コラム一覧

1 ｜ 研究開発の内容

本プロポーザルが掲げる「次世代AIモデルの研究開発」は、超大規模深層学習ベースの現在の人工知能（AI）が抱える問題点（資源効率、実世界操作、論理性、信頼性、安全性など）を克服する新しいAIモデルを創出し、人・AI共生社会を実現するための研究開発を意味する。

深層学習（Deep Learning）を用いたAIが急速な発展を示し、特に生成AI（Generative AI）、基盤モデル（Foundation Model）、大規模言語モデル（Large Language Model：LLM）などと呼ばれる超大規模深層学習ベースのAIが、大きな社会インパクトをもたらしつつある。OpenAIやGoogleなどの米国ビッグテック企業が技術開発で大きく先行し、国内ではその後追い開発や応用開発が活発化している。上述の問題点を含むAIリスクに対して、喫緊の問題への対策・ルール整備なども国際的な議論の中で進められている。

このような状況を踏まえ、本プロポーザルが狙うのは、その先の次世代AIモデルを創出する基礎研究[1]の戦略強化である。そのための基本的な考え方（方針）は以下のA～Dである。

【方針A】次世代AIモデルへの発展を、AI単体の機能・性能の発展だけでなく、AIと他者（人間および他のAI）や社会との関係の発展にまで広げて考える。なぜならば、現在のAIモデルは、資源効率、実世界操作、論理性、信頼性、安全性などに問題を抱えている一方、驚異的な性能を示しているがゆえに人間の活動や社会の仕組みにも変化をもたらしつつある。現在の問題に対する改善だけでなく、AIと人間・社会の関係の在り方からも、次世代AIモデルの要件を考えていくことが必要になる。

【方針B】次世代AIモデルに向けた多様なアプローチを認めつつ、それらの間で問題意識や知見の共有を促す。技術ブレークスルーの可能性を幅広に探りながらも、ばらばらな取り組みとせず、シナジーや融合を促すことが、基礎研究を加速することになるからである。

【方針C】日本が抱える社会課題の解決や日本の特性に即した社会発展を支えるためのAI技術は自国で保有すべきである。他国のビッグテック企業と競争するのではなく、わが国として経済安全保障を確保しつつ、労働人口減少に直面する日本社会の発展や人々のWell-beingのために、現在のAIモデルの限界・問題点を克服することが必要である。

【方針D】ビッグサイエンス化などAI研究の形態が変化し、基礎研究においてもビッグテック企業が大きく先行していることや、総合知による取り組みが不可欠になっていることを踏まえた、基礎研究推進の新しい体制づくりや研究エコシステムづくりを考える。

この方針のもと、次世代AIのための基礎研究として取り組むべき研究開発課題を以下の4点とした。

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
② AI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
③ AI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
④ AIと人間・社会の関係とその在り方に関する研究

1 ここでの「基礎研究」とは、目標とする機能や概念の要件を定義し、その要件に関わる現象のメカニズムを解明したり、要件を達成する基本原理を生み出したりする取り組みを意味する。情報分野においては、基本原理が生み出されたならば、それをソフトウェアとして実装することは比較的短期間で可能であるが、通常さらに、実用的な性能を達成するための改良や、社会や人々に受容されるようにするための改良が必要になる。そのような実装・改良の結果、社会や人々の受け止め方に応じて、目標とする要件を再定義する必要が生じ、「基礎研究」にフィードバックされることも、短いサイクルの中でしばしば起きる。なお、基礎研究の位置付けについては、文部科学省による分類[2)]を参考にした。

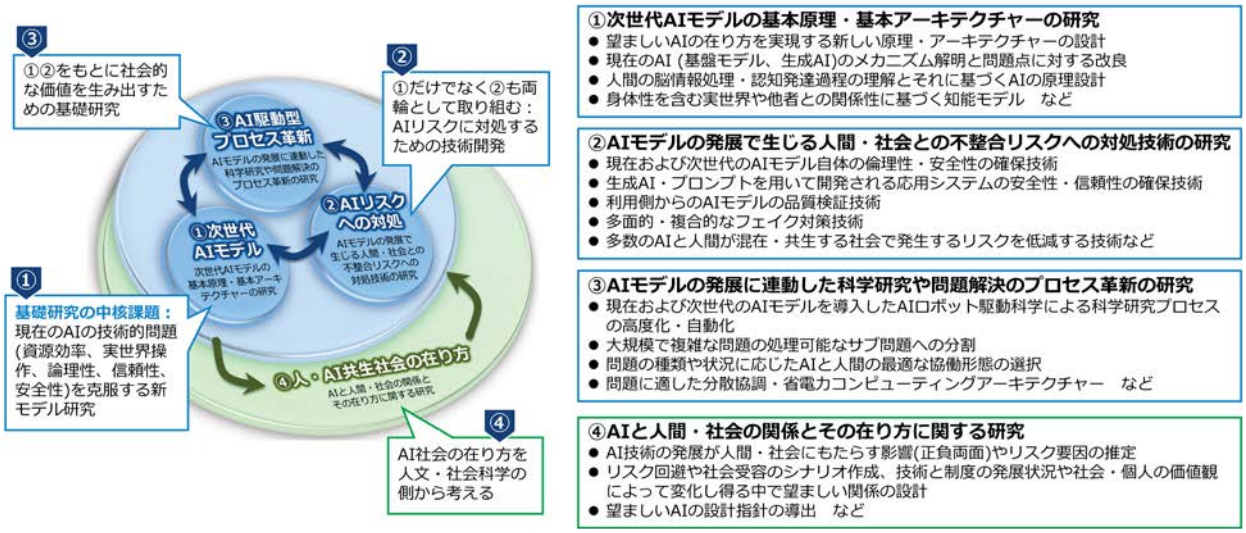


図 1–1　　取り組むべき研究開発課題

①が次世代AIモデル設計の中核になるが、②のAIリスクへの対処技術にも同時に取り組むことが不可欠である。①と②はいわば車の両輪のような関係となる基礎研究だが、そこから社会的な価値を生み出すための基礎研究が③のプロセス革新研究である。また、①②③は主に情報科学技術の研究であるのに対して、④は人・AI共生社会の在り方に関する人文・社会科学の研究である。④によって①②③の技術要件・指針が定まる一方、①②③の技術発展の結果として④の考え方や目標が変わり得る。これら四つは、相互に影響し合い、連動しながら進展するものである。これらの関係と①～④の詳細を図 1–1 に示す。

この取り組みを推進するにあたり、方針Dで触れた研究開発形態の変化を踏まえる。すなわち、AI分野の研究開発は基礎研究においても、ビッグサイエンス化、ハイスピード化・ハイインパクト化、非オープン化の傾向が強まっており、この状況下での研究開発体制・基盤やプログラムの在り方を考え、それを支える研究エコシステムをつくることが必要である。その中には、図 1–2 に示すa～hなどの要素が含まれる。

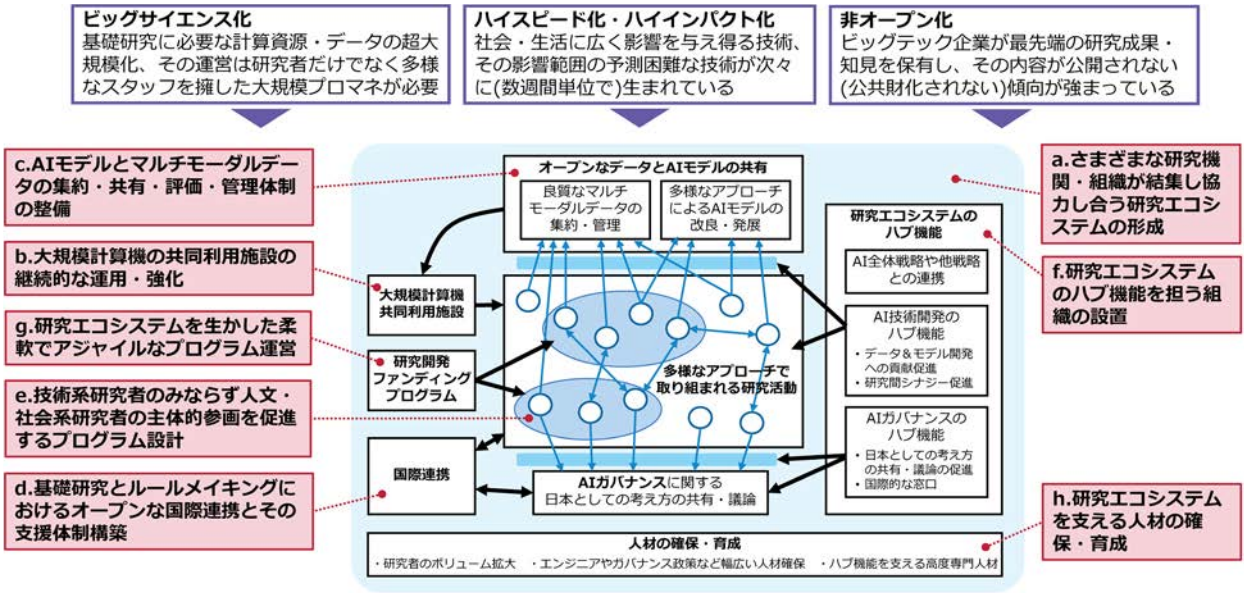


図 1–2　　研究エコシステムの概念と推進方策

2 ｜ 研究開発を実施する意義

2.1 現状認識および問題点

本節では、まず現状認識として、基盤モデル・生成AIの登場のインパクトを簡単に述べるとともに、基盤モデルと生成AIの定義と関係を説明し（2.1.1）、その開発状況（2.1.2）とリスク面（2.1.3）の認識、および、関連する政策の動向（2.1.4）についてまとめる。次に、基盤モデル・生成AIの課題について広く全体像を示した上で、現状認識を踏まえて、本プロポーザルで扱う「次世代AIモデルに関わる基礎研究」の対象範囲を示し（2.1.5）、現在の取り組み状況と問題認識について述べる（2.1.6）。

2.1.1 基盤モデル・生成AIのインパクト

AI技術の発展が著しい。2022年11月末に公開されたChatGPTは爆発的にユーザー数を拡大し、わずか2カ月で1億人のアクティブユーザーを獲得した[1]。ChatGPTに代表されるような、入力された自然言語に対して応答するAIは、生成AI（Generative AI）と呼ばれている。言語の他に画像・映像・音声・動作などマルチモーダルの応答や、言語以外のモーダルを併用した入力も可能になりつつある。まるで人間のような自然な会話や、専門的な知識・能力を備えているかのような応答が可能になり、ChatGPTは米国の名門大学MBA（経営学修士）や医師資格試験に合格するレベルであるとか、米国の司法試験で上位10%に入るレベルであるといった報告もなされている[3), 4), 5)]。

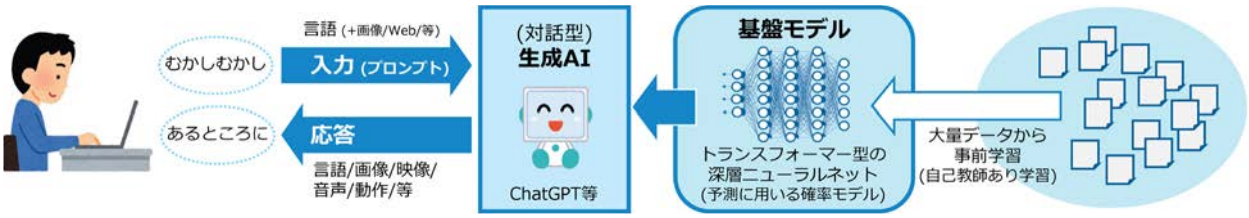


図2–1　生成AIと基盤モデル

現在の生成AIは、超大規模深層学習によって作られた確率モデルに基づいて、与えられた入力文（プロンプトと呼ばれる）の続きを予測することで応答を生成する（図2–1）。この確率モデルは、大量データからさまざまな関係性を事前学習した超巨大な深層ニューラルネット構造を成し、基盤モデル（Foundation Model）や大規模言語モデル（Large Language Model：LLM）と呼ばれる。基盤モデルは、Stanford Institute for Human-Centered Artificial Intelligenceによって命名され、「大量かつ多様なデータで訓練され多様な下流タスクに適応できるモデル」と定義されている[6)]。言語を主としたものはLLMと呼ばれることが多い。本稿では、超大規模深層学習によって作られた「基盤モデル」を、現時点で最先端のAIモデルを意味する用語として主に扱い、その機能面に焦点を当てるときには「生成AI」という用語を用いる。

1　当時、史上最速といわれた。ただし、その後、2023年7月にThreadsが5日で1億人を達成した。

従来のAIは、タスクごとに教師データを用意して学習させるタスク特化型のAIモデルだったが、基盤モデルは一つのモデルでさまざまなタスクに適用でき、汎用性の高さが革新的である。また、生成AIは、AIに行わせたいタスクの内容説明や指示を自然言語で与えることができるので、その利用に際して特段高いスキルが要求されないことから、幅広い層に利用が一気に広がった。

このような技術・利用の両面での飛躍により、基盤モデル・生成AIは大きな社会インパクトを生みつつある。人間の知的作業全般に急速な変革をもたらし、産業、研究開発、教育、創作などさまざまな分野に幅広く波及すると見込まれている。既にさまざまな応用開発が進んでおり、表2–1はその一例である。 McKinsey & Companyによると[1)]、生成AIは世界経済に年間数兆ドル相当の価値をもたらす可能性がある。わが国では特に、労働人口減少に対する生産性向上や産業・経済の活性化につながるとの期待が大きい。

表2–1　基盤モデル・生成AIの応用例

分野	応用例
自然言語処理	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、音声テキスト変換など
コンピュータービジョン	Text-to-Image生成、Image-to-Text生成、画像分類、物体検出、ビデオ生成、キャラクター生成など
ソフトウェアエンジニアリング	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、文書化など
インターフェイス	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、実行計画など
科学・教育・医療などの応用	創薬、ゲノム解読、医療診断、個人教師など

2.1.2 基盤モデルの開発状況

2020年7月にOpenAIから発表されたGPT-3は、モデルのパラメーター数が1750億個と、それまで主流モデルだったBERT（Googleが2018年10月に発表）の500倍以上という超大規模モデルであり、それ以降、モデルの超大規模化が急速に進んだ[2]。例えば、2021年10月にMicrosoftとNVIDIAが発表したMT-NLG（Megatron-Turing Natural Language Generation）のパラメーター数は5300億個、2022年4月にGoogleが発表したPaLM（Pathways Language Model）のパラメーター数は5400億個である。その後、GPT-3の改良版であるGPT-3.5をベースに初代のChatGPTが開発され、2022年11月末に公開された。2023年3月にはChatGPTの基盤モデルとしてより強化されたGPT-4が追加された。 Googleもこれに対抗してBardを開発し、試験的公開を経て2023年5月に一般公開に至った。 Bardのベースとなる基盤モデルは、当初LaMDAが用いられたが、一般公開時にはPaLM2が用いられた。なお、OpenAIのGPT-4、GoogleのPaLM2といった最新の基盤モデルは、それ以前のモデルからさらに規模拡大・強化されているが、モデル規模を含め、詳細な技術情報は非公開とされている。

モデル発展の系譜を示す図2–2からも分かるように、最先端の基盤モデル・生成AIの技術開発では、OpenAI、Google、Metaなどの米国ビッグテック企業が大きく先行している。一方、AI分野の研究開発・

2　基盤モデルやLLMの詳細な動向は、既発行の報告書[7)]の中にまとめている。また、LLMに関するよく知られたサーベイ論文[8), 9)]がある。図2–2は同論文[9)]から引用した。なお、モデルの規模は通常、モデルのパラメーター数で表される。モデルによっては、規模を変えた複数タイプを発表しているものもある。その場合、本稿におけるモデル規模の記載は、複数タイプのうちで最大規模のもので代表させている。

ビジネスで米国と並んで2強となった中国でも、基盤モデル・生成AIの開発が進められている。その代表的なものを挙げるならば、北京智源人工智能研究院（BAAI）の悟道（WuDao 2.0）はパラメーター数が1.75兆個、百度（Baidu）の文心一言（ERNIE 3.0 Zeus）や阿里巴巴（Alibaba）の通義千問（Tongyi Qianwen 2.0）もパラメーター数が数千億個の規模といわれている。しかし、最先端モデルとして話題になるのはOpenAIやGoogleなど米国ビッグテック企業のものであり、モデル規模では米国ビッグテック企業のものを上回るものが開発されているものの、現状、中国製モデルは性能面で米国の最先端モデルほど注目されるには至っていないようである。

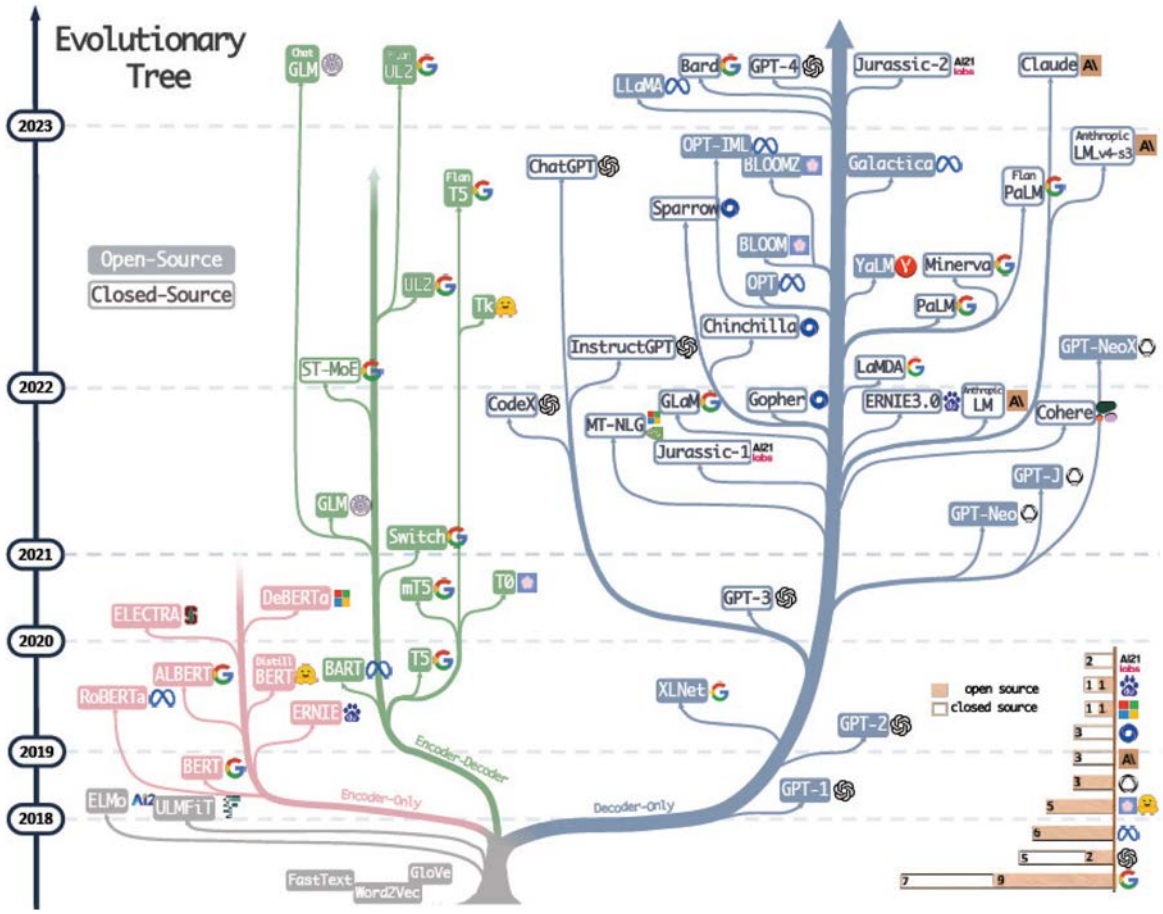


図2–2　基盤モデル（LLM）の系譜[9)]

国産基盤モデルの開発状況は、中規模モデルの後追い開発となっている（図2–3）。ある程度の規模を持つモデルの開発に比較的早い段階で着手したのは、LINE（現LINEヤフー）と韓国NAVERの共同開発プロジェクトである（HyperCLOVA、2020年11月に発表、2022年には820億パラメーター規模のモデルを開発）。2022年11月末にChatGPTが登場して以降、2023年には多数の企業から中規模モデルの事業化あるいは開発計画の発表が相次いだ（CyberAgent、NEC、NTT、Preferred Networks、ソフトバンク他）。このような多数参入の背景には、MetaのLlama（2023年2月公開）、Llama2（2023年7月公開）に代表されるオープンソースモデルの広がりがある。加えて、前述したような産業・ビジネスの成長領域としての期待がある。日本の人々がChatGPTの活用に積極的であることは、openai.comドメインへの国別アクセスシェアが、米国、インドに次いで世界3位[10)]であることから分かる。

また、産業・ビジネス用途だけでなく研究開発のために、大学や国の研究機関においても国産基盤モデル

開発が立ち上がっている。特に国立情報学研究所（NII）を中心としたLLM-jp（LLM勉強会）[3]は、大学・企業などから700名以上が参加し、オープンで日本語に強いモデル開発や原理解明に取り組んでおり、2023年10月に130億パラメーター規模のモデルを公開した。他にも、情報通信研究機構（NICT）は同年7月に400億パラメーター規模、東京大学の松尾豊研究室は同年8月に100億パラメーター規模、東京工業大学の岡崎直観研究室・横田理央研究室と産業技術総合研究所は共同で同年12月に700億パラメーター規模などのモデルを公開した。これらについて、さらに大規模なモデル開発も進められている。

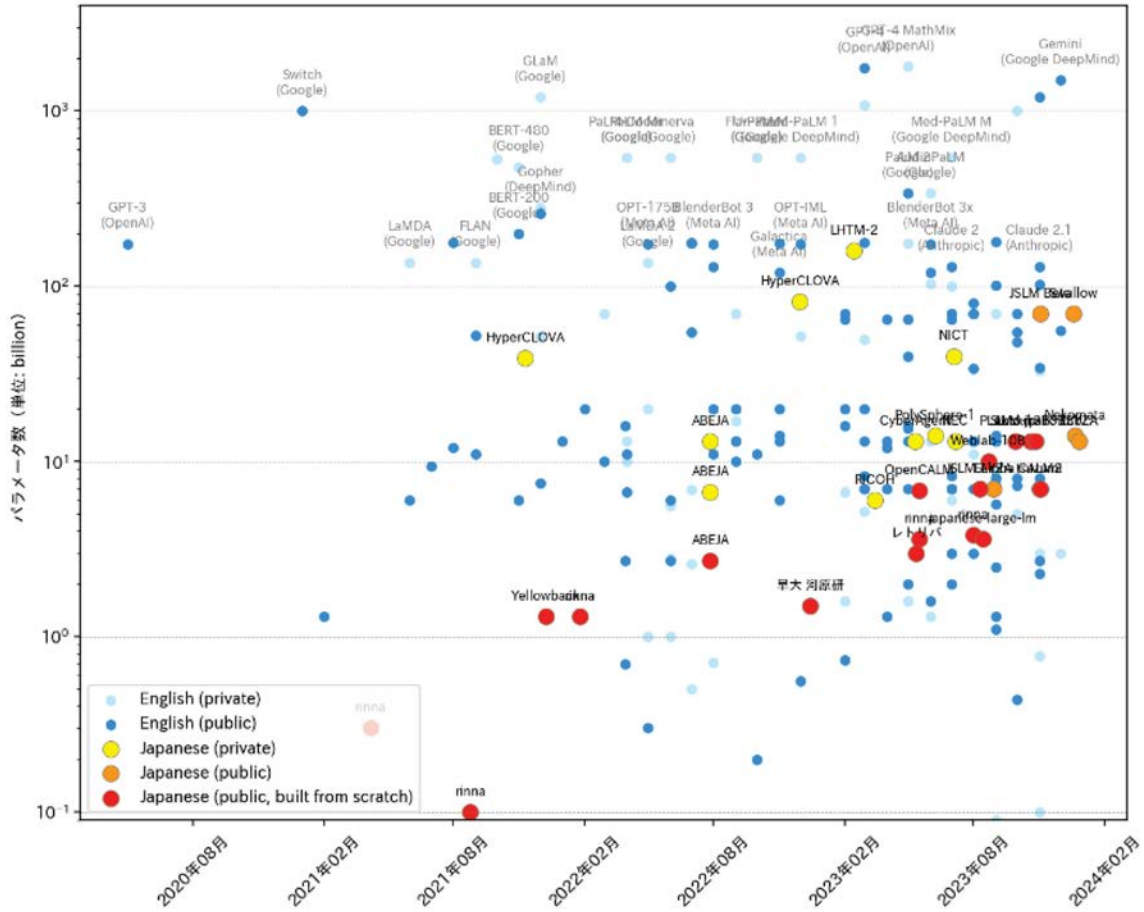


図2–3　　国産基盤モデル（日本語LLM）の開発状況

（出典　https://github.com/llm-jp/awesome-japanese-llm　2024年1月1日時点）

日本国内の基盤モデル開発力の底上げのため、経済産業省は2024年2月に、プロジェクトGENIAC（Generative AI Accelerator Challenge）を立ち上げた。GENIACでは、計算資源の提供、利活用企業やデータホルダーとのマッチング支援、グローバルテック企業との連携支援やコミュニティーイベントの開催、開発される基盤モデルの性能評価などが計画されている。第1期採択事業者として、ABEJA、Preferred Elements（Preferred Networksの子会社）、東京大学（松尾豊教授ら）、Sakana AI、ストックマーク、情報・システム研究機構（国立情報学研究所）、Turingが選ばれた。

3　2024年4月には国立情報学研究所内に大規模言語モデル研究開発センター（仮称）が設置される予定である。

コラム1

生成AIブームが生まれた技術的要因

2022年から2023年にかけて、生成AIの利用が爆発的に拡大した。このようなブームが生まれたのは、Web上で使える形で提供されたことも大きな要因の一つと考えられるが、技術的な面では、以下に示す3点が要因として挙げられる。

1点目は、予測精度の劇的向上である。第3次AIブームの初期、画像認識・音声認識などのパターン認識に深層学習が適用され、従来法よりも大幅な精度改善が示されたが、自然言語処理への適用では従来法から芳しい精度改善が得られていなかった。しかし、意味の分散表現、アテンション機構を用いたトランスフォーマーモデル、自己教師あり学習などの技術が次々に開発・導入され、さらに超大規模化が進められたことで、最近5年間の自然言語処理の劇的な精度向上に結びついた[7]。その超大規模化による精度向上については、モデルの予測精度が、計算リソース、データセット規模、モデル規模という3変数のべき乗則に従うというスケーリング則（Scaling Laws）[11] が観測されたことで（図2–4）、超大規模化に拍車がかかった[4]。すなわち、大量データを学習させてモデルの規模を巨大化するほど予測精度が向上する。

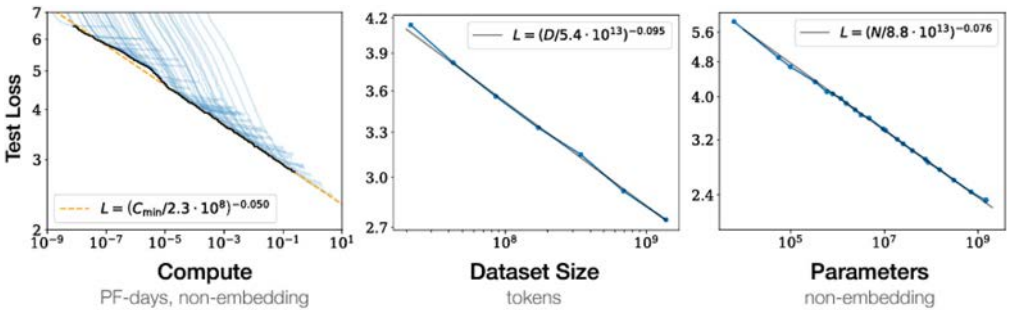


図2–4　スケーリング則[11]

2点目は、対話型ユーザーインターフェイスの採用である。実は、1点目として挙げた予測精度の劇的向上は、AI研究者コミュニティーにおいてGPT–3の登場時点（2020年）で大きな話題になっていた。これが2022年11月末のChatGPT登場によって、一般にも爆発的に利用が広がったのは、自然言語での対話（チャット）という、一般ユーザーに分かりやすく使いやすいインターフェイスが採用されたことが、大きな要因になったと考えられる。

4　スケーリング則に加えて、モデル規模がある点を超えると性能が急激に向上するという創発的能力（Emergent Abilities）が見られたという報告[12] がある。ただし、この観測結果は、測定する指標の設定の仕方によるなどの疑問が示されている[13]。

3点目は、人間の意図・価値観に合わせてAIを振る舞わせる仕組み（いわゆるAIアライメント）にも取り組まれたことである。チャットボット関連の事件として、2016年3月23日に、MicrosoftのTayがTwitter上で公開されたが、悪意を持ったユーザーとの対話によって、ごく短時間で差別主義的思想に染まってしまい、開始後16時間で強制終了となったことがよく知られている。このような問題を回避し、対話型生成AIがより適切な応答を返せるようにするため、以下のような複数通りの対策が取られている。

- 学習データ選別：不適切な内容のデータは取り除き、学習に使わない。
- RLHF（Reinforcement Learning from Human Feedback）：人間のフィードバックを用いた強化学習によって基盤モデルをチューニングする。ChatGPTでは、（1）プロンプトと望ましい応答のペアを追加学習、（2）応答文のスコア付けモデル（真実性＋無害性＋有益性）を人間による比較評価から学習、（3）学習したスコア付けモデルを用いて、大量のプロンプトで強化学習を行い、基盤モデルをチューニング、という手順によってRLHFを実現している[14)]。
- Content Moderation：不適切な表現（性的/暴力/ヘイトなど）を含む応答を検出してブロックする。

2.1.3 深刻化するAIリスク

基盤モデル・生成AIへの期待の一方で、急速に発展するAIが生むリスク、人間・社会の価値観や既存システムとの不整合によるELSI（倫理的・法的・社会的課題）の懸念が高まっている[5), 15), 16)]。その代表的なものを図2–5に示す。

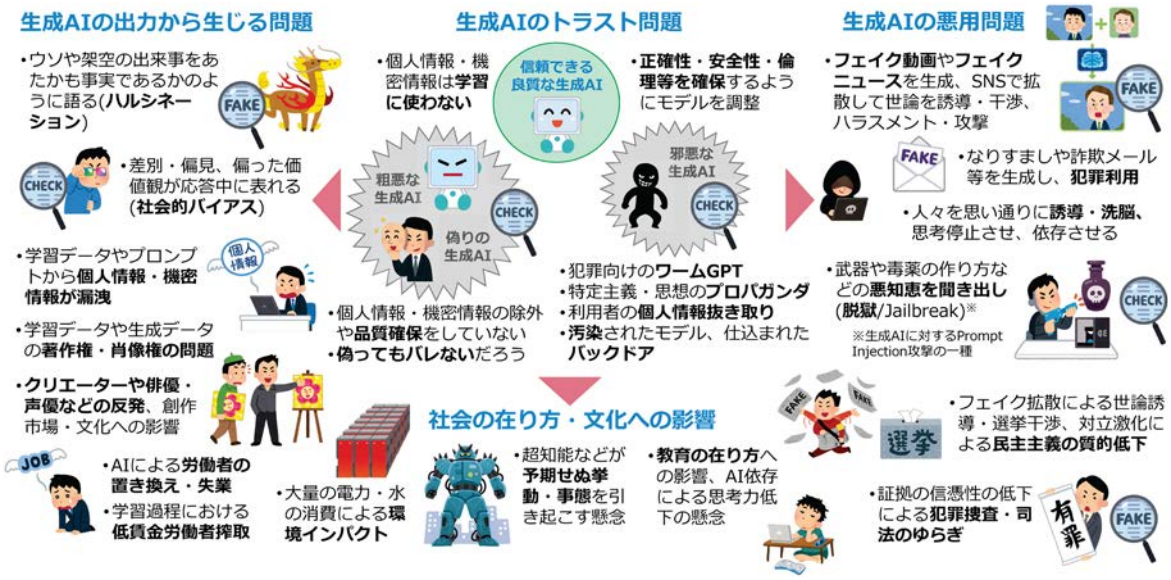


図2–5　生成AIがもたらすリスク

特によく指摘されるのは、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（Hallucination）」と呼ばれる問題である。生成AIは自然でもっともらしい応答を返してくるものの、確率モデルに基づいて続きを予測しているだけで、意味を理解したり、論理推論を行ったりしてはいないために生じる現象である。

生成AIの出力から生じる問題としては他にも、差別や偏見、偏った価値観が応答中に表れる社会的バイアスの問題や、学習データやプロンプトから個人情報・機密情報が漏洩するリスク、学習データや生成データに関わる著作権・肖像権の問題なども顕在化している。

また、生成AIが悪用される問題も深刻化しつつある。最も社会問題化しているのはフェイク問題である。生成AIを用いることで、人にはもはや見破ることが困難なほどのフェイク動画やフェイクニュースを簡単に生成できてしまう。これをSNSで拡散して世論を誘導・干渉したり、ハラスメントや新種のサイバー攻撃に用いられたりといったことによる被害やネガティブインパクトが拡大しつつある。なりすましに使ったり、詐欺メールなどを生成したりして犯罪に利用されることも起きている。

生成AIは、社会にとって望ましくない応答（性的なもの、暴力的なもの、ヘイトなど）は出力しないように調整されているが、そのようなガードをかいくぐって、武器や毒薬の作り方のような悪知恵を聞き出す「脱獄（Jailbreak）」と呼ばれる行為も問題になっている。

一方、生成AI自体も、信頼できる良質な生成AIばかりとは限らない。良質な生成AIは、正確性・安全性・倫理性などを確保するようにモデルが調整されているし、個人情報や機密情報は学習に使わないように配慮されている。しかし、生成AIのオープンソース版が公開されたり、カスタマイズされたGPT（My GPTs）が簡単に作れたりするようになり、生成AIが乱立するようになると、その中には、邪悪な生成AI、粗悪な生成AI、偽りの生成AIなども混じるようになると予想される。犯罪向けのワームGPTが既に存在するといわれるし、特定主義・思想のプロパガンダを意図した生成AIや、ユーザーの個人情報を抜き取ろうとする生成AIなどの懸念も生じる。

さらに、社会の在り方・文化に対する影響についても、必ずしも良い影響・効果だけでなく、悪影響の懸念も指摘されている。フェイク問題に起因するものとしては、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・司法のゆらぎなどが懸念される。また、人並みの能力を示すことから、AIによる労働者の置き換え・失業、教育の在り方への影響やAI依存による思考力低下なども懸念されている。大量の電力・水の消費による環境インパクトの懸念や、AIが急速に進化することで超知能などが予期せぬ挙動・事態を引き起こす懸念なども生じている。

なお、超知能などが予期せぬ挙動・事態を引き起こす懸念として、欧米ではAIが人類を破滅させる事態を招くという危機意識が強い（一方、日本ではこの意識は弱い）。この事例として、2023年3月に米国FLI（Future of Life Institute）から出されたオープンレター「強力なAIシステム開発の一時停止を求める」[5]に、Yoshua Bengio、Stuart Russell、Elon Musk、Steve Wozniak、Yuval Noah Harariなどの著名人を含む3万人以上が署名している。また、イエールCEOサミットでの調査において、119名のCEOのうちの42%が「AIは今から5～10年以内に人類を破滅させる潜在的可能性がある」と回答した[6]。

また、AI ELSIとは別の観点として、基盤モデル・生成AIは社会・生活のさまざまな場面・活動に用いられる汎用技術になり得ることから、先行するビッグテック企業提供サービスへの過度な依存は、日本にとって経済安全保障面や科学研究・産業の国際競争力の面でのリスクとなる。

5 https://futureoflife.org/open-letter/pause-giant-ai-experiments/（2024年1月1日時点で33,709名が署名）

6 https://www.cnn.co.jp/tech/35205269.html（2023年6月14日のCNN Exclusiveの記事）

2.1.4 関連政策の動向

以上のような期待と懸念に対して、2023年には、イノベーションの推進策とリスクに対するルール作りの両面から、各国で生成AI関連の政策立案が急速に進んだ。

AIの開発・運用における安全性管理・統治の枠組み（AIガバナンス）について、これまでの動向を振り返ると、日本政府の「人間中心のAI社会原則」をはじめ、2019年にAI社会原則に関する国・国際レベルでの議論が活発化した。国際的な原則として、経済協力開発機構（Organization for Economic Co-operation and Development：OECD）が、2019年5月に「人工知能に関するOECD原則」（OECD Principles on AI）をまとめ、これに42カ国が署名した。ユネスコ（国際連合教育科学文化機関、United Nations Educational, Scientific and Cultural Organization：UNESCO）での検討も2019年から始まり、

表2–2　米中欧日のAI政策関連トピックと特徴

	AI政策関連トピック	特徴・特記事項
米国	2017年1月 アシロマAI原則 2019年3月 IEEE 倫理的に配慮されたデザインEAD1e 2022年10月 AI権利章典のための青写真 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OpenAI、Google、Microsoftなど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	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。
中国	2017年7月 次世代人工知能発展計画（AI2030） 2019年6月 次世代AIガバナンス原則 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	国際学会でも躍進著しく、米中2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。
欧州	2018年4月 AI for Europe 2019年4月 信頼できるAIのための倫理指針（欧州委員会） 2020年2月 AI白書（欧州委員会） 2021年4月 AI法案（欧州委員会） 2023年6月 生成AI対応を盛り込みAI法案修正（欧州委員会） 2023年7月 AI条約の統合版ドラフト公開（欧州評議会） 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	各国のAI戦略に加え、研究・イノベーションの枠組みプログラムによる欧州内連携の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順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。
日本	2017年7月 国際的な議論のためのAI開発ガイドライン案（総務省） 2018年8月 AI利活用原則案（総務省） 2019年3月 人間中心のAI社会原則（内閣府） 2019年6月 AI戦略2019（その後、2021、2022に更新） 2019年8月 AI利活用ガイドライン（総務省） 2021年8月 AI原則実践のためのガバナンスガイドライン（経済産業省、2022年1月に更新） 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、AIに関する暫定的な論点整理 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 2024年1月 AI事業者ガイドライン案（総務省、経済産業省）	「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研究をけん引する体制を整備。 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。

2021年11月に「AI倫理勧告（first draft of the Recommendation on the Ethics of Artificial Intelligence）」が全193加盟国によって採択された。

その後、原則から実践へとフェーズが移行し、2023年からは基盤モデル・生成AIに関わるルール作りが重要な論点となっている。表2–2に米中欧日のAI政策関連トピックとその特徴をまとめた。2023年の基盤モデル・生成AIに関わる政策トピックに注目すると、まず2021年4月に公表された欧州のAI法案（AI Act）に関して、生成AIに対する規制も盛り込むことが2023年6月に可決された。具体的には、基盤モデルのプロバイダーに、違法コンテンツの生成の防止、著作権法で保護されたコンテンツを学習した際の公表、リスク評価の実施など、透明性への取り組みを義務付ける内容になった。また、日本はG7議長国として、急速な発展と普及が国際社会全体の重要な課題となっている生成AIについて議論する広島AIプロセスを2023年5月に立ち上げた。2023年9月の中間閣僚級会合、10月のマルチステークホルダーハイレベル会合を経て、12月の閣僚級会合で安全安心・信頼できる高度なAIシステムの普及を目的とした指針と行動規範から成る初の国際的政策枠組み「広島AIプロセス包括的政策枠組み」を取りまとめ、G7首脳に承認された。もう一つのトピックは、11月1日・2日に英国主催で開催されたAI安全サミットである。英米欧中日を含む28カ国が支持した「ブレッチリー宣言」では、AI社会原則などで主に訴求されてきたAI倫理よりむしろ、予期せぬ事態への危機感、偽情報対策、AI安全への国際協力などが強調された。

AIガバナンス関連政策や2023年の生成AI規制動向についてまず述べたが、AI関連政策は米中欧日それぞれの立ち位置がある。詳しくは表2–2の右カラムにまとめたが、米国は、基盤モデル・生成AIの研究開発で世界をリードしており、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。中国は、政府の後押しもあってAIを活用した産業に勢いがあることや、政府が監視・管理社会の構築のためにAIを活用していることも特徴的である。欧州は、AIに関わる国際ルール作りを通して米国・中国およびビッグテック企業に対抗しており、EUルールをグローバル企業が順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」も生まれている[17)]。日本は、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（その詳しい内容を図2–6に示す）、生成AIに関わる取り組みも進めており、AIガバナンスでは、アジャイルガバナンス、ソフトローを指向している。

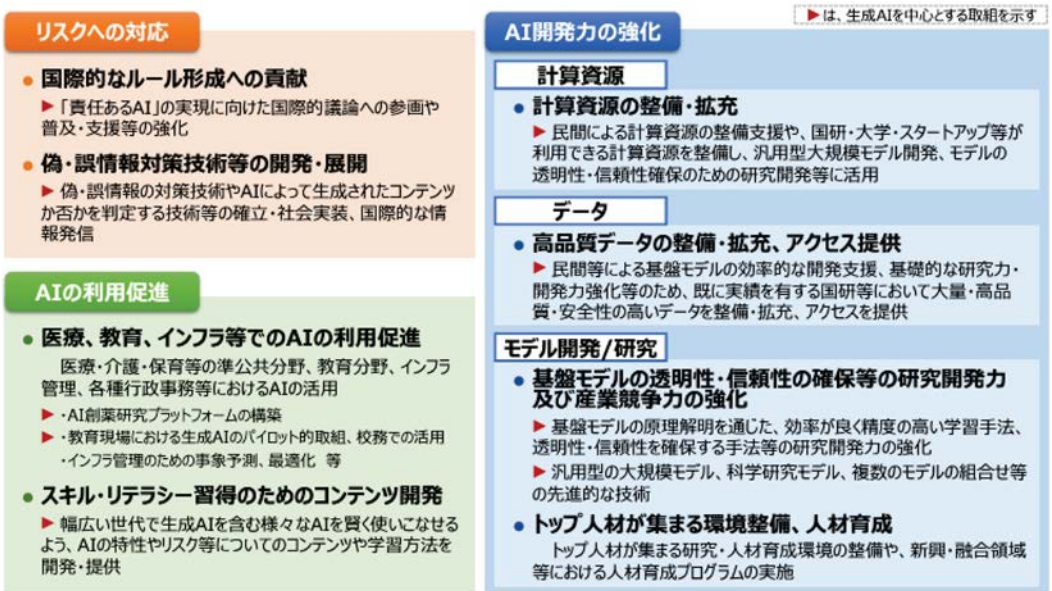


図2–6　わが国のAI関連の主要施策

内閣府AI戦略会議第4回（2023年8月4日）の資料に示された案

2.1.5 基盤モデル・生成AIの課題の全体像と戦略提言の位置付け

ここまで技術開発面、リスク面、政策面から、基盤モデル・生成AIについての現状認識をまとめた。これを踏まえて、基盤モデル・生成AIに関わる課題の全体像を示し、本プロポーザルで扱う対象範囲を明確化する。

まず、図2–7に基盤モデル・生成AIに関わる課題の全体像をまとめた。この図では、左が実務寄りで右が学術寄り、下が共通基盤で上が応用個別として、八つの課題を配置した。下段中央を出発点とした時計回り順で、それらを簡単に説明する。

図2–7　基盤モデル・生成AIに関わる課題の全体像

下段中央の「基盤モデル構築」は、大規模な基盤モデルの実装、そのための計算機環境の構築や高速処理技術、データ収集・選別・整備などである。下段左の「基盤モデル運用」は、新しいデータを追加してモデルを更新するプロセスがベースとなるが、継続運用が可能になるようなビジネスモデルや、基盤モデルに特定の思想を仕込むなどの懸念のない運用体制のトラスト（信頼）確保も含む。その上、中段左の「基盤モデル周辺拡張」は、現状の基盤モデルが不得手な機能（最新情報検索、数式処理、論理推論など）をプラグインなど外部連携によって補ったり、それらを用いたワークフローを自動設計・最適化したりするものである。

上段は利活用・応用実現の層である。上段左の「基盤モデル応用開発（API利用）」は、既存の基盤モデルのAPIを利用した応用開発であり、前述のように、既にさまざまな応用が開発されている。上段中央と上段右にまたがる「分野固有基盤モデル開発・活用」は、既存の基盤モデルをそのまま使うのではなく、分野固有のデータを用いて、ファインチューニングしたり、別途、分野固有のモデル（ニューラルネット）を作ったりして分野の用途向けのシステムを開発するものである。

その下、中段右の「AIリスク対処研究」は、基盤モデルなどのAI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に関する基礎研究である。中段中央の「利活用時の問題対処の仕組み」は、生成AIの入出力データに関わる問題として、著作権や肖像権への対処、生成AIが出力したものか/人間が作成したものかの区別などを含む。下段右の「次世代AIモデル研究」は、現在の基盤モデルからさらに発展させるための基本原理・基本アーキテクチャーに関する基礎研究である。基盤モデルの高効率化、生成AIの高性能化、基盤モデルのメカニズム解明、人間の知能からヒントを得た新しい知能モデルの探求、そのような新モデルに適したコンピューティングなどを含む。

これまで述べてきたように、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である。このような現

状認識を踏まえると、課題の全体像（図2–7）についての取り組み状況は、図2–8に示すように、左3分の2は「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」であり、右3分の1が「基礎研究として重点的に取り組むべき課題」だと考える。わが国の政策においても、図2–6に示した通り、前者・後者ともに手が打たれつつあるが、喫緊の課題として動きが活発なのは前者であり、中長期的な目線での戦略立案が必要な後者については、まだ検討が限定的だと思われる。

本プロポーザルが対象とするのは後者であり、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。

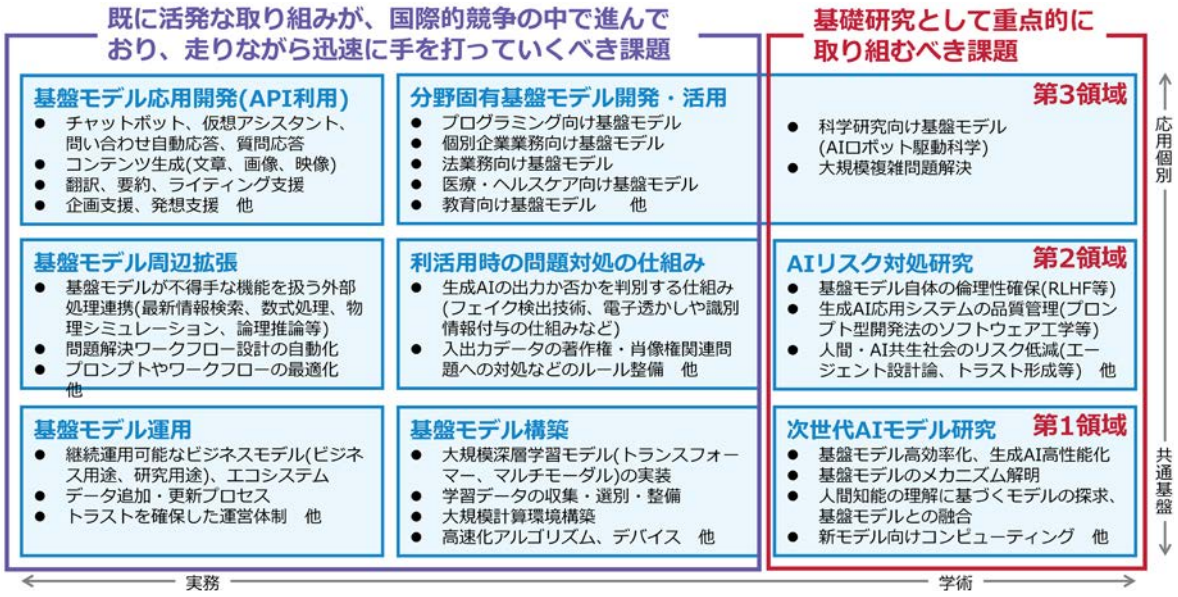


図2–8　課題の全体像に対する現状認識

2.1.6 次世代AIモデルに関わる基礎研究の状況と問題認識

基盤モデル・生成AIに関わる課題の全体像（図2–7、図2–8）の右3分の1に該当する三つの基礎研究領域について、現在の取り組み状況と、克服すべき問題が何かを述べる。なお、現在の取り組み状況・事例のより詳しい情報は付録Bで取り上げる。

（1）第1領域「次世代AIモデル研究」の取り組み状況・問題認識

現在の基盤モデルからさらに発展させるための基本原理・基本アーキテクチャーに関する基礎研究である。

スケーリング則を生かした超大規模深層学習に基づく現在（2023年時）の基盤モデル・生成AIは、それ以前の深層学習モデルが目的特化型AIであったのに対して、高い汎用性とマルチモーダル性を実現するなど、大きな発展を遂げた。これは衝撃的な成果で、非常に強力な道具になることは間違いないが、その仕組みは、2.1.1で述べたように、大量データからの学習で作られた確率モデルに基づいて、入力文の続きを予測するものだということから来る問題・限界がある。

この問題・限界はさまざまな捉え方があるが[18]、上記のような仕組みを採っている結果、人間の知能に及ばないという現象が見られるものとして、主に以下の5点が挙げられる。

●**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）を必要とすること。

超大規模深層学習は膨大な学習データと計算資源を用いる。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り（図2–9）、電力消費面も懸念されている。人間の脳は20ワット程度で動

いているといわれることから、電力効率には改善の余地があると考えられる。

●**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・行動が苦手であること。
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、実世界状況の動的な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なこと。
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。また、大きなタスクの解決方法は、さまざまな要因や取り組み方（分解と組み合わせ）があり得て、簡単に確率モデルに落とし込むことができない。

●**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こと。
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。また、確率モデルに基づくので、どうしても不安定性は残り、どのような条件でどのような結果が得られるかを100%予測したり保証したりすることはできない。

●**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こと。
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。人間の知能と比べたときに現在のAIには自発性が欠けているが、そもそも将来のAIに自発性を持たせることが良いのかについては議論が必要であろう。そこで、「自発性」は、第1領域の「克服すべき問題」として扱わないことにする。

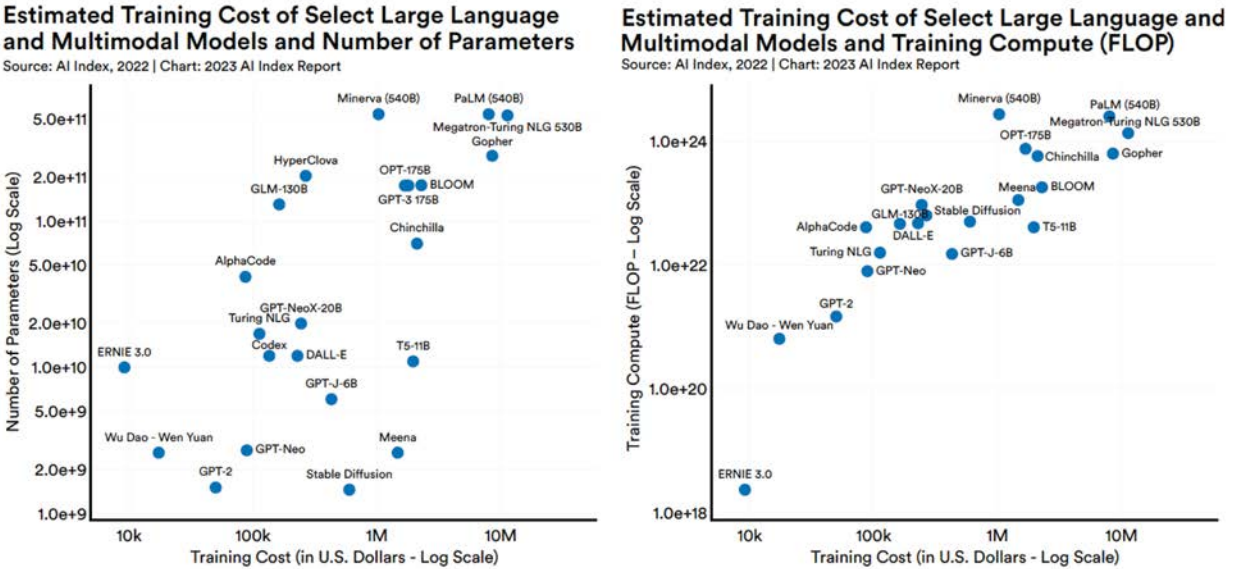


図2–9　基盤モデルの学習に必要な費用の見積もり[19)]

そこで、資源効率、実世界操作（身体性）、論理性、信頼性、安全性という五つを、克服すべき問題と捉え、現在の取り組み状況を大きく4通りのアプローチに分けて説明する。それらの具体的な研究事例は**付録B.1**に示す。また、各アプローチによる五つの問題の克服の見込みを**表2–3**にまとめた。

表2–3　現在の基盤モデルの問題に対するアプローチの例

アプローチ	資源効率	実世界操作（身体性）	論理性	信頼性	安全性
（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	×	△	△	△	△
（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	?	?	?	?	?
（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	○	○	○	△	○
（d）知能を人間・社会との関係性の側面から発展させる	?	△	△	○	○

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

アプローチ（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

現在、実用に近いところで活発に取り組まれているのは、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タイプのアプローチである。

その一つは、基盤モデルの苦手な処理を外部モジュールとして用意し、基盤モデルと連携させる方法である。例えば、基盤モデルが苦手な数式処理、物理シミュレーション、最新情報検索などが外部モジュール（プラグイン）として用意され、基盤モデルと組み合わせて使えるようになっている。

もう一つは、学習に用いるデータやプロンプトを工夫する方法である。例えば、テキストや画像だけでなく、ロボット動作データを学習させることで、ある程度の実世界操作に対応できるように拡張する取り組みがある。

信頼性・安全性や論理性についても、基盤モデル（生成AI）の出力の良し悪しを人間が判定してフィードバックする強化学習（**コラム1**で触れたRLHFなど）も使われている。

アプローチ（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

（a）のような基盤モデルの外側に処理・手順を足していくアプローチで対処できる問題範囲は限定的である。さらに、外側に処理・手順を足していくことで、処理負荷や必要なデータ量が増加し、資源効率はますます悪化してしまう。したがって、外付け改良の積み上げではなく、基本原理・基本アーキテクチャーに立ち返ったアイデア創出が、根本的な対処には必要であろう。

これにつながるものと考えられるのが、現在の基盤モデルのメカニズム解明の取り組みである。現在の基盤モデルの仕組みで、なぜこれほど賢く見える応答が得られるのか、その一方で、どのようにしてハルシネーション（もっともらしくウソを返す現象）が起きるのかなど詳しいメカニズムが明らかになっていない。そのようなメカニズムの解明は、問題の克服のために基本原理・基本アーキテクチャーをどのように見直すと良いかの示唆につながる。ただし、現時点ではメカニズムの解明が進んでいないので、どの問題をどう克服できるのかは未知である。

アプローチ（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

問題を克服するための基本原理・基本アーキテクチャーのヒントになるのは、人間の知能のメカニズムである。人間の知能のメカニズムの全容はまだ解明されていないが、近年の脳科学の進展で得られた研究成果や認知科学・発達科学・心理学・行動経済学の知見などが、問題克服の有益なヒントになり得る。これまでも人間の知能に関する知見は、深層学習や強化学習をはじめとしてAIのモデル・原理に示唆を与えてきた。さまざまな研究成果・知見が活用され得るが、既にAIモデルとしての試作が進められているものの代表例として、

二重過程理論や予測符号化理論がある。

二重過程理論は、人間の思考は、経験に基づいた即応的な思考を担うシステム１と、ある種抽象化されたモデル・知識を参照した熟考的な思考を担うシステム２で構成されるというモデルである（表2–4）。このような２タイプの情報処理は、人間の脳においても異なる箇所で担われていると考えられている。深層学習・強化学習による帰納型のパターン処理はシステム１で実行される処理に相当するが、システム２で実行される演繹型の記号推論処理は十分にカバーされていない。それ以前の深層学習モデルから基盤モデルに発展したことで、システム２も部分的にカバーされたように思われるが、メカニズムの詳細は明らかになっておらず、論理構築・論理推論が苦手なことから、システム２の実現には至っていない。帰納型の処理だけで精度を高めようとすると、どうしても大量データからのボトムアップ学習が必要になり、資源効率の問題は回避できない。演繹型の推論を用いるならば、必要なデータのみ取りに行くというトップダウン制御が可能になり、資源効率の問題に対処できようになる。また、論理性の問題にも対処できるとともに、トップダウンの制御によって安全性の問題にも対処しやすくなる。

表2–4　　二重過程理論における２タイプの思考

システム１：即応的な思考	システム２：熟考的な思考
高速	低速
直感的	論理的
無意識的	意識的
非言語的	言語的
習慣的	計画・推論型

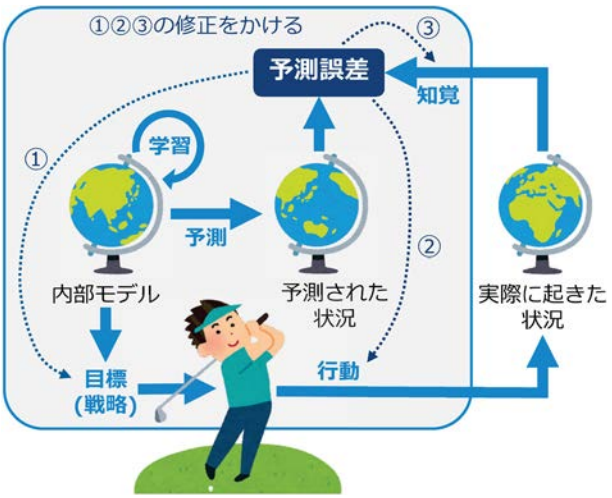


図2–10　　予測符号化理論の概念

一方、予測符号化理論は、実世界操作（身体性）の問題に対処できると期待される。これは、発達科学や認知発達ロボティクスの分野で研究開発が進んでいるもので、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・推論などの認知機能を発達させていく過程をモデル化しようというものである。予測誤差最小化原理（自由エネルギー原理）によって、さまざまな認知発達を統一的に説明することが試みられている。すなわち、現時刻・空間の信号から、将来や未知空間の信号を予

測できるように、その対応関係（内部モデル）を学習するメカニズムであり、身体や環境からの感覚信号と、脳が内部モデルをもとにトップダウンに予測する感覚信号との誤差を最小化するように、内部モデルを更新したり、環境に働きかけるような運動を実行したりする（図2–10）。この原理を応用したスマートロボットやシミュレーターなども開発されている。

アプローチ（d）知能を人間・社会との関係性の側面から発展させる

信頼性の問題に対しては、知能を人間・社会との関係性の側面から発展させる取り組みが注目される。上で述べてきたようなアプローチは、AI単独の発展、つまり、1個のAIとしてできることが拡大・高度化していくことを指向している。しかし、AI単独で発展・高度化しても、AIのブラックボックス性のため、信頼性は必ずしも高まらない。この問題に対処する技術の一つが説明可能AI（XAI）であるが、さらに踏み込んだ技術チャレンジとして、コモングラウンドの実現が対話システム研究分野において検討されている。コモングラウンドとは、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バックグラウンドのことである。現在の生成AIは、一見、相手のことを分かっているかのような応答を返すが、実際は、事前学習した確率モデルや、プロンプトとして与えた情報からのIn-context Learningを用いて、ある意味反射的に予測応答を返しているだけである。生成AIと相手（ユーザー）との間にコモングラウンドは形成されていない。AIの人間・社会との関係性に着目したアプローチは、第2領域「AIリスク対処研究」にもつながり、安全性の問題にも効果がある。マルチエージェント社会のメカニズムデザインなど、関連するアプローチについては第2領域「AIリスク対処研究」にて触れる。

以上、現在の基盤モデルの問題への対処について、4通りのアプローチの取り組み状況を述べた。表2–3に示すように、資源効率、実世界操作（身体性）、論理性、信頼性、安全性という、克服すべき五つの問題に対して、部分的な対処は検討されているものの、根本的な解決につながる次世代AIモデルの基本原理・基本アーキテクチャーの創出が望まれる。

（2）第2領域「AIリスク対処研究」の取り組み状況・問題認識

基盤モデルなどのAI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に関する基礎研究である。ここで克服すべき問題は、2.1.3の図2–5に示した、生成AIがもたらすさまざまなリスクである。第1領域で問題の一つとして挙げた安全性（および信頼性）の問題とも重なる。

さまざまなリスクに対して、2.1.4「関連政策の動向」で示したように、喫緊の問題に対する制度や規範を策定する動きが、国内の政策検討でも、国際的なルール形成の場でも進められている。しかし、制度策定は一定の効果があるものの、制度や規範だけでAIリスクを完全に抑え込むことは無理である。制度策定とともに、制度を強化したり補完したりする技術的な対策を組み合わせることが必要である。また、技術開発の進展に応じて、必要な制度も変わる。

この第2領域は、生成AIのリスクへの技術的な対処を対象とする。生成AIの開発・利用においてリスクが発生し得る具体的な5通りのケースを図2–11に図示した。以下、それぞれのケースの概要と、技術的な取り組み状況を説明する。なお、ケース1・2・3は開発段階、ケース4・5は利用段階で発生する。

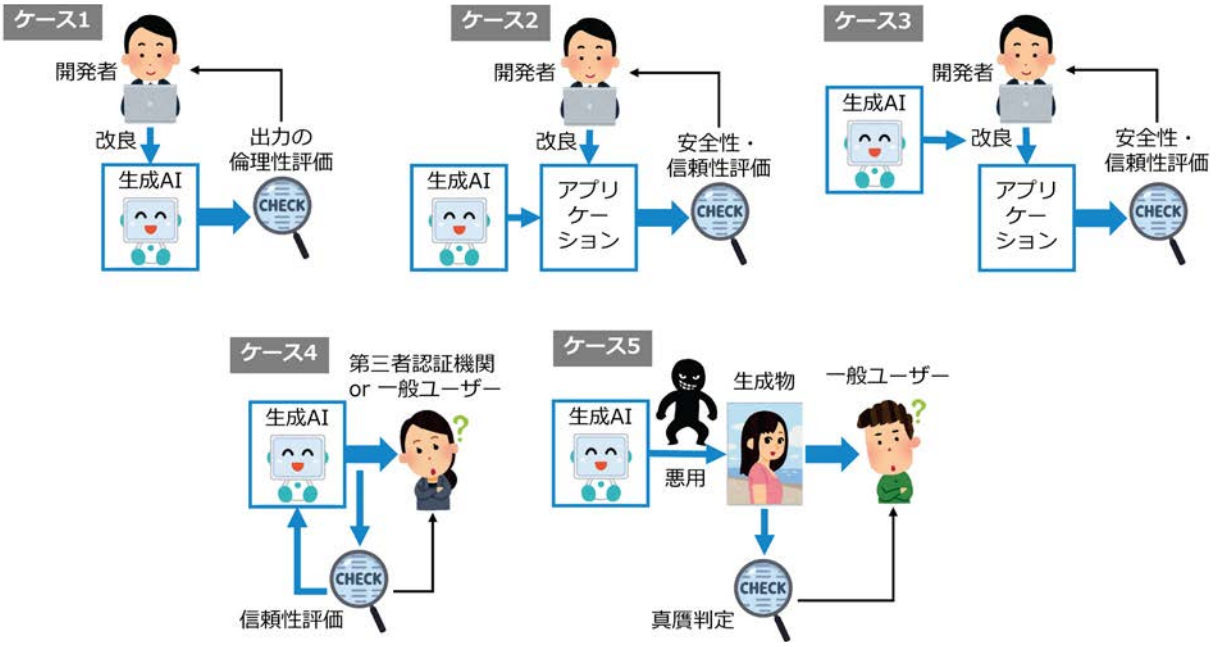


図2–11　生成AIの開発・利用におけるリスクが発生するケースと技術的対処

- **ケース1：**

生成AIが倫理的に不適切な出力をしてしまう問題に対処するための技術開発である。ハルシネーション、差別的・暴力的な応答、著作権・肖像権を侵害するような出力などを回避することが求められる。**コラム1**に示したように、現状、開発時の対処として、真実性・無害性・有益性の面から人間によるフィードバックを加えるRLHF、データ選別、結果フィルターなどの方法が試みられている。

- **ケース2：**

生成AIを組み込んだアプリケーションを開発するケースにおいて、開発するアプリケーションの安全性・信頼性を確保するための技術開発である。生成AIをAPIで呼び出すという形が主であるが、OpenAIが提供し始めたMy GPTsのように、Web上で生成AIを特定用途に簡単にカスタマイズして公開するサービスも使えるようになった。正常系の処理を簡単に作ることができても、異常系の処理や予期していない状況での振る舞いに十分対策できているかは不安が残る。また、生成AI特有の脆弱性（プロンプトインジェクションなど）が攻撃されるリスクも高い。簡単にアプリケーションを開発できて、開発者の裾野が広がることで、リスクを持ったシステムが乱立してしまう恐れもある。2018年頃からソフトウェア工学と機械学習が交わる研究分野・研究コミュニティーが立ち上がり、機械学習応用システム開発の品質管理のための方法論や技術群が開発されてきたが、生成AIアプリケーションの品質管理は新しい問題で、2024年に入って、研究コミュニティー内でこの問題が議論されるようになってきたところである。

- **ケース3：**

アプリケーション開発作業（プログラミング、テスト、デバッグなど）を生成AIに支援させるケースにおいて、開発するアプリケーションの安全性・信頼性を確保するための技術開発である。例えば、自然言語で指示して、生成AIにプログラムコードを生成させたり、テストプログラムを生成させたり、デバッグを代行させたりといったことが実際に行われつつある。ケース3はケース2と同じくソフトウェア工学の範疇なので、懸念事項や取り組み状況はケース2と同様である。

• **ケース4：**
このケースはユーザー視点の問題を扱う。2.1.4でも述べたように、多数の生成AIやその応用システムが乱立してきたときに、信頼できる良質な生成AIと、邪悪な生成AI、粗悪な生成AI、偽りの生成AIなどを判別できるかという問題がある。制度面からは第三者認証制度を設ける手があるが、第三者認証機関が審査するにせよ、ユーザー自身が直接判定するにせよ、生成AIが信頼できるものであるかを外部から評価する手法の開発が望まれる。制度面の動きはあるが、技術開発面はこれからと思われる。

• **ケース5：**
生成AIをフェイク生成やなりすましなどに悪用するケースに対して、それを見破るための技術開発である。最近数年で取り組みが急速に活発化している。フェイクメディアの検出技術は、メディア（動画・画像・音声など）の詳細解析、機械学習などによって、不自然さを判定するといったことが行われている。フェイクテキストの検出技術は、内容、表現スタイル、拡散パターン、情報源などをもとに判定することが試みられている。ただし、フェイク生成方法と検出技術はいたちごっこであるとともに、生成AIは急速に高精度・高品質化して、もはや人間の目で見破ることは無理と思われるレベルに達している。そこで、異なるアプローチとして、電子透かしを埋め込んだり、ブロックチェーンなどで管理したり、出所や流通・改ざん経路をたどれる仕組みも立ち上がりつつある。

以上、生成AIの開発・利用においてリスクが発生し得る具体的な5通りのケースについて、AIリスクへの対処技術への取り組み状況を述べた。ケース1とケース5は最近数年で研究開発が活発化しているが、ケース2・3・4は問題が徐々に認識されつつある段階で、取り組みが立ち上がるのはこれからである。いずれについても、リスクを完全に除去できるものではなく、総合的・体系的な取り組みを進めることで、リスク軽減を図ることが必要であろう。

（3）第3領域「分野固有基盤モデル開発・活用」のための基礎研究の取り組み状況・問題認識

「分野固有基盤モデル開発・活用」領域は、例えばプログラミング向け基盤モデルや個別企業業務向け基盤モデルなどのように、実務に適用されて効果を示しつつ改良が進められているものから、科学研究向け基盤モデルのように、パラダイムを大きく変える可能性を持つ基礎的な取り組みまで幅広い。ここで第3領域として扱うのは、基礎研究としての取り組みが必要とされる後者である。

特に重要性が高まっているトピックとして、AIロボット駆動科学を取り上げ、その取り組みの概要と、科学研究向け基盤モデルの適用状況について述べる[7), 21)]。具体的な研究開発事例は付録B.3にまとめた。

AIロボット駆動科学（図2–12）は、「知識→（仮説推論）→仮説→（演繹）→予測→（検証実験）→実験結果→（帰納）→知識」というサイクルの繰り返しによって、検証済みの仮説を蓄積・洗練していく科学研究のサイクル[7]について、AIを用いた大規模・網羅的な仮説生成・探索による人間の認知限界・バイアスを超えた科学的発見と、ロボットを用いた仮説評価・検証のハイスループット化を実現しようという取り組みである。囲碁の世界において、AIを用いたAlphaGoが世界トップクラスの棋士に圧勝した。そこでAlphaGoが行っていた膨大な可能性の探索から導出された打ち手は、人間の棋士には思いもよらなかった手を含んでいたが、それはその後、新手として人間の棋士も取り入れるようになった。同様のことは、今後、科学的発見においても起こり得る。このような科学研究サイクルの強化・自動化は、科学研究の国際競争力、さらには

7 この図2–12は、研究者個人や個々の研究グループでの活動サイクルである。ここで得られた説が研究コミュニティーによって受け入れられるまでには、このサイクルの外側にもう一つの研究コミュニティーによって検証・評価されるというサイクルを経ることになる。

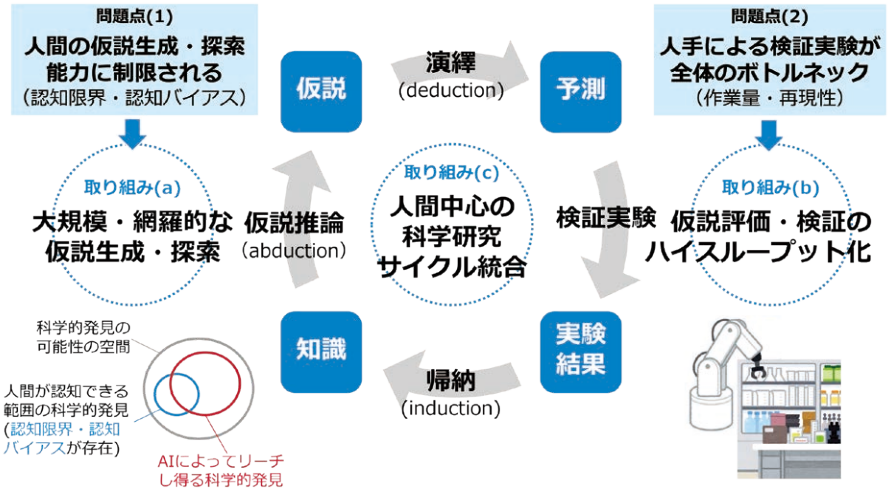


図2–12　AIロボット駆動科学

表2–5　AIロボット駆動科学における三つの取り組み

（a）大規模・網羅的な仮説生成・探索	（b）仮説評価・検証のハイスループット化	（c）人間中心の科学研究サイクル統合
仮説の大規模生成 ・不完全な情報での仮説推論 ・因果グラフの推定・探索 ・仮説空間の探索の高効率化 **既存知識との整合性検証** ・高次元の中間表現の解釈 ・整合・矛盾の迅速評価 ・制約条件や境界条件の設定	**ハイスループット・高精度実験** ・実験ロボットの高度化 ・ハイスループット計測 ・ネットワーク化、標準化 **実験計画の自動化** ・実験計画の生成・最適化 ・バーチャルスクリーニング ・ベイズ推定、能動的観測	**アーキテクチャー設計** ・プラットフォームへの統合 ・知識体系の集積化 ・人間との相互作用の設計 **科学の科学** ・科学研究サイクルの理解 ・セレンディピティーの実装 ・「発見」「理解」の科学

さまざまな産業の国際競争力を左右する。

AIロボット駆動科学では、図2–12および表2–5に示すように、主に三つの取り組みが進められている。

（a）大規模・網羅的な仮説生成・探索：

超多次元の現象（非常に多くのパラメーターで記述される現象）から規則性を見いだすことは人間には困難だが、深層学習を用いれば、それが可能になりつつある。複数の異なる専門分野の知識をつなぎ合わせた推論による、仮説の生成・探索は人間には困難だが、今後、論理推論の枠組みを分野横断で実行できれば、それが可能になるかもしれない。

（b）仮説評価・検証のハイスループット化：

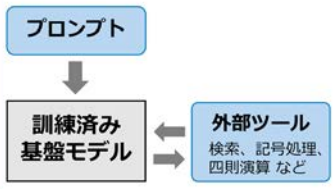
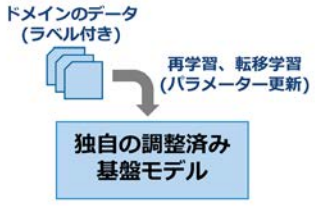
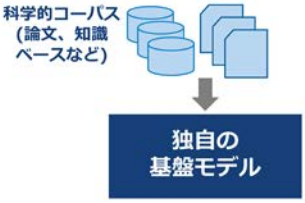
ロボットなどによる物理的な実験の自動化技術も含め、科学的発見プロセスを構成するさまざまな技術を一つのプラットフォーム上に統合する取り組みが進められている。そこでは、計算量や物理的操作を抑える、効率の良い処理フローや絞り込みアルゴリズムが必要になる。

（c）人間中心の科学研究サイクル統合：

（a）と（b）を統合して、仮説の生成と検証のクローズドサイクルを実行する人間参加型のシステムとして実現するものである。まずは生命科学分野や材料科学分野で限定した条件のもとで取り組みが立ち上がり、科学研究サイクルの中にAIやロボットを組み込んで、サイクルを回す事例が少しずつ生まれている。その具体的な事例は付録B.3に示す。

そこに基盤モデル・生成AIが登場し、AIロボット駆動科学においても、その活用も始まりつつある。その導入・応用のパターンは、大きく分けると表2–6のような3種類（A・B・C）がある。

表2-6　基盤モデル・生成AIの科学研究応用のパターン

	(A)そのまま利用する	(B)分野に適応させる	(C)独自モデルを開発する
モデル	調整済みモデルをそのまま利用 (パラメーター固定)	モデルを修正(層の追加・削減、 再学習でパラメーターを更新)	独自モデルをスクラッチから 事前学習し、再学習もする
適応策	プロンプトエンジニアリング、 外部ツール連携	再学習、転移学習、ファインチューニング、蒸留、RLHF など (基本的には教師あり学習／強化学習)	
データ	不要(学習しない)	再学習に必要	大量・多様なデータが必要
計算資源	ホストするには必要	再学習に必要	事前学習には大規模なものが必要
構成			
用途例	研究サイクル内の一般業務・ 周辺業務の効率化など	科学文献の概要・動向把握、 言語・画像系作業の効率化など	分野固有データ(タンパク質、 元素・化合物他)からの予測など
システム例	ChatGPT、Bard、 Walframプラグインなど	Galactica、PMC-LLaMA、 MatSciBERTなど	RGN2、HelixFold、ESMFold、 EMBER3Dなど

（A）は既存の調整済みの基盤モデル（ChatGPT・Bardなど）をそのまま利用するというパターンである。基盤モデル自体では足りない機能（検索・四則演算など）は外部ツール連携によって補完し、基本的にはプロンプトによる指示・手順を工夫することで、科学研究に役立つようなタスクを実行させる。必ずしも科学分野固有の詳細知識がなくとも、一般業務や周辺業務の手順を言語で説明して自動化できることで、効率化が図れる。

（B）は既存の基盤モデルをもとに、対象分野のデータを追加学習させて、その分野に適応させるというパターンである。その分野固有の詳細知識を持たせることができるので、その分野の文献の概要・動向把握や、テキストからの情報抽出・関連付けや画像の認識・判定などの精度が高まる。専門性の高い作業の効率化・自動化を進めることができる。

（C）は既存の基盤モデルを流用するのではなく、独自モデルを開発して利用するというパターンである。対象分野データを大量に集めて、スクラッチから学習して独自モデルを作る。既存の基盤モデルのベースとなっている言語モデルは、単語系列が文章を構成し、その間の関係性を捉えたものであることから、例えば、生命科学分野ではアミノ酸系列がタンパク質を構成し、化学・材料分野では元素・結合が分子を構成するといった類似性に着目して、アミノ酸系列や元素・結合を大量にトランスフォーマーに学習させることで、その分野固有の予測問題を扱うことが可能になる。

このような基盤モデル・生成AIの適用によって、AIロボット駆動科学における三つの取り組み（a、b、c）のそれぞれにおいて、次のような効果が見込まれている。まず、（a）大規模・網羅的な仮説生成・探索においては、生成AIによる予測機能が仮説生成の可能性や効率を高めると期待される。また、文献などの知識からの関連情報抽出・整理や整合性検証などにも活用できる。（b）仮説評価・検証のハイスループット化においては、画像・映像・音などで得られる観測結果の分析・検証の効率を高めると期待される。（c）人間中心の科学研究サイクル統合においては、対話型のインターフェイスが、科学研究のプロセスにおける手順の対話的指示や結果の対話的理解に役立つ。

しかし、第1領域で挙げた現在の基盤モデル・生成AIの問題点は、AIロボット駆動科学においてもやはり問題になる。特に論理性の問題は、仮説生成・探索の精度を高めるために、取り組むことが不可欠である。また、実世界操作（身体性）の問題は、検証実験に用いるロボットの柔軟で正確な制御のために重要な課題である。

2.2 社会・経済的効果

2.1節で述べたように、基盤モデル・生成AIは、人間の知的作業全般に急速な変革をもたらし、産業、研究開発、教育、創作などさまざまな分野に幅広く波及すると見込まれている。 McKinsey & Companyによると[1)]、生成AIは世界経済に年間数兆ドル相当の価値をもたらす可能性がある。わが国にとっては、労働人口減少に対する生産性向上や産業・経済の活性化につながるという期待も大きい。

表2–7には、生成AIが社会のさまざまな側面でもたらす変化と、そこから考えられる「望ましい未来」と「望ましくない未来」を示した。「望ましくない未来」を回避し、「望ましい未来」を実現するための政策上の着眼点には、強い基礎・基盤を生み出す研究開発、人材育成、データ構築・整備体制、活用を促進しながらリスクを回避するプロアクティブなルール整備、海外巨大企業による独占・寡占ではない日本のエコシステムづくりなどが考えられる。本プロポーザルでは、特に「強い基礎・基盤を生み出す研究開発」の切り口から戦略提言を行うことで、表2–7における「望ましい未来」として挙げたような価値実現に貢献する。

表2–7　生成AIが社会にもたらす変化と政策上の着眼点

生成AIによってもたらされつつある変化		望ましい未来	望ましくない未来
産業	• 様々な産業において知的作業が効率化・自動化	• 生産性が高まり、産業が成長・活性化	• 活用できる人材不足で産業低迷 • 活用が広がっても収益を上げるのは海外企業
科学研究	• 科学的発見の新しい道具となり、人間の限界を超えた可能性探索、研究開発のスループット加速	• 科学的発見の新しい道具として活用先行、研究力・技術開発の国際競争力が向上	• 活用に出遅れ、研究力・技術開発の国際競争力が低下 • 知見やデータが海外サーバに流出
個人生活	• 個人の創造性や可能性を拡大する道具・アシスタントとして活用が進む	• 誰も取り残さず社会参画促進 • やりたいことに集中でき、皆が活躍できる機会拡大	• プライバシー、個人の行動情報を海外企業が掌握 • 活用差によって格差が拡大
社会の安全性	• 犯罪や事故を未然回避 • 社会的意思決定のために幅広い情報を集約・活用	• よりよい社会的意思決定がなされて、社会の安定、安心安全が進む	• フェイク生成がサイバー攻撃手段化 • 真偽不明が招く社会混乱、法の揺らぎ
文化・思想	• 基盤モデル運営の中に国の文化やノウハウが反映されていく	• 日本の文化やノウハウを集積・発展、世界に発信	• 海外企業にノウハウが集積 • 日本が他国の文化に染まる
実現を支える技術発展の方向性 • AIモデルの扱うタスクの高度化 • マルチモーダル化、実世界操作、分散協調 • セキュア、高速化、エコ、小型・軽量化 等		**望ましくない未来を回避し、望ましい未来を実現する政策上の着眼点** • 強い基礎・基盤を生み出す研究開発、人材育成、データ構築・整備体制 • 活用を促進しながら、リスクを回避するプロアクティブなルール整備 • 海外巨大企業による独占・寡占ではない、日本のエコシステム作り	

2.1.6では、現在の基盤モデル・生成AIには、資源効率、実世界操作（身体性）、論理性、信頼性、安全性に関する問題があることや、2.1.3や図2–5に示したようなさまざまなリスクがあることを述べた。本プロポーザルに示す研究開発によって、これらの問題やリスクが克服・軽減されることで、例えば以下のようなことが可能になり、上述した「望ましい未来」の実現につながる。

- 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。

2.3 科学技術上の効果

AIモデルの発展を図2–13に図示した。ここでは、第1次ブーム期の探索ベースのAIを第1世代AI、第2次ブーム期のルールベースのAIを第2世代AI、第3次AIブーム期の機械学習ベース（深層学習ベース）のAIを第3世代AIと呼ぶ。現在の生成AIブームは第4次AIブームともいわれているが、AIモデルとしては第3世代AIのアーキテクチャーを超大規模化したものなので、基盤モデルは第3.5世代AIと呼ぶことにする。

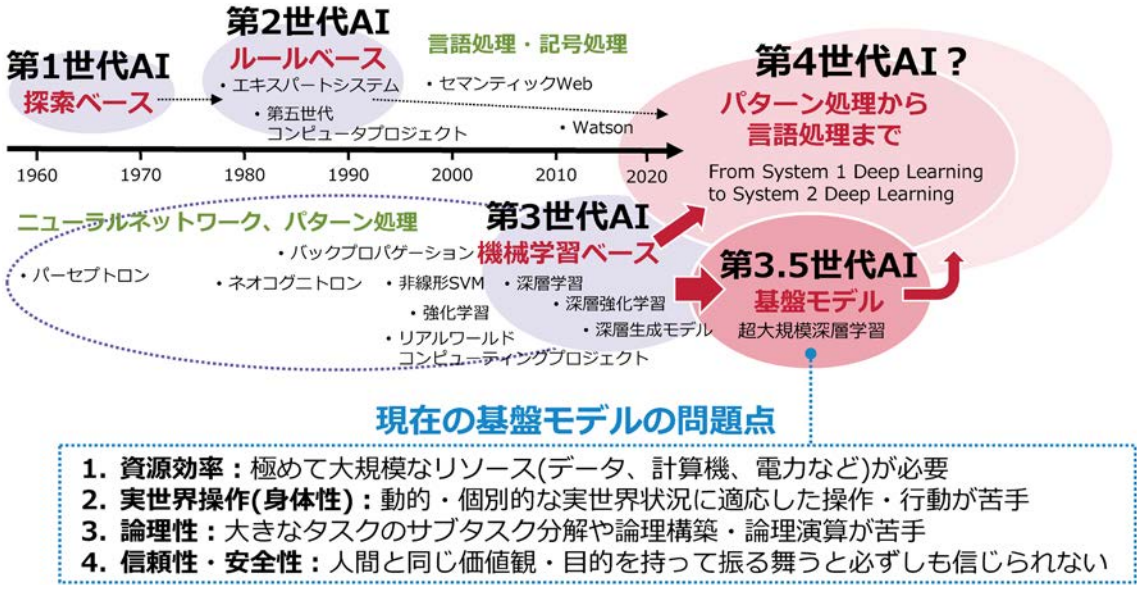


図2–13　AIモデルの発展

この流れの中に位置付けると、本プロポーザルで示した次世代AIモデルは第4世代AIである。本プロポーザルは、この新しい研究開発テーマ領域を生み出し、その取り組みを加速・拡大する。現在（2023年時）、基盤モデル・生成AIは大きなインパクトをもたらし、産業界・アカデミアともそれ一辺倒になっている感がある。基盤モデル・生成AI自体の研究開発は重要不可欠だが、それだけでなく、次世代AIモデルという観点からの取り組みも推進しておくことが、次のフェーズで先行するチャンスを生む。これが基盤モデル・生成AI自体の改良に対しても新たな示唆を生み得る。

なお、CRDSでは、2020年3月に戦略プロポーザル「第4世代AIの研究開発―深層学習と知識・記号推論の融合―」[20] を公開している。この中で描いた第4世代AIは、深層学習（第3世代AI）に知識・記号推論を融合させて、パターン処理から言語処理までを統一的な枠組みで実現しようというものであった。帰納型AIと演繹型AIの融合、即応的AI（システム1）と熟考的AI（システム2）の融合という方向性を示した。実際にこのような方向性の研究開発が進みつつあった一方、基盤モデル（第3.5世代AI）が驚異的な性能を示した。そこで、基盤モデルを含む急速なAI技術発展の状況や多数の専門家の意見なども踏まえて、次世代AIモデル（第4世代AI）へのアプローチについて、より多面的な切り口から、より広い可能性を見据えて検討したのが本プロポーザルである。図2–13において、「第4世代AI？」という表記を用い、二重の円で表現したのは、このような意図による（二重円の内側の円は2020年時点、外側の円は今回のより広い可能性を見据えて拡大したもの）。

また、AI分野の研究開発は基礎研究であってもビッグサイエンス化やハイスピード化・ハイインパクト化が

進んでいる状況において、**第4章**にて後述する推進方法（研究エコシステムなど）を含めて考えるならば、情報系研究者と人文・社会系研究者と産業界が協働した総合知による研究開発や、その中での多様なスキルを持つ人材育成も促進される。

3 ｜ 具体的な研究開発課題

繰り返し述べたように、本プロポーザルは、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を提言するものである。そのため、**第2章**に述べた現状認識や克服すべき問題認識に基づき、**第1章**に列挙した四つの研究開発課題を設定した。

四つの研究開発課題は、次の①～④である。

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［**次世代AIモデル**］

② AI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［**AIリスクへの対処**］

③ AI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［**AI駆動型プロセス革新**］

④ AIと人間・社会の関係とその在り方に関する研究［**人・AI共生社会の在り方**］

これら四つの研究開発課題は、**2.1.1**に示した関連動向の俯瞰的調査、および、50名を超える専門家へのインタビュー、ワークショップや学会などでの意見交換・議論を通した有望な方向性の見極めなどを通して定めた。その際の基本的な考え方（方針A～D）は**第1章**で述べた。

これら四つの研究開発課題の内容と課題間の関係は**図3–1**の通りである。

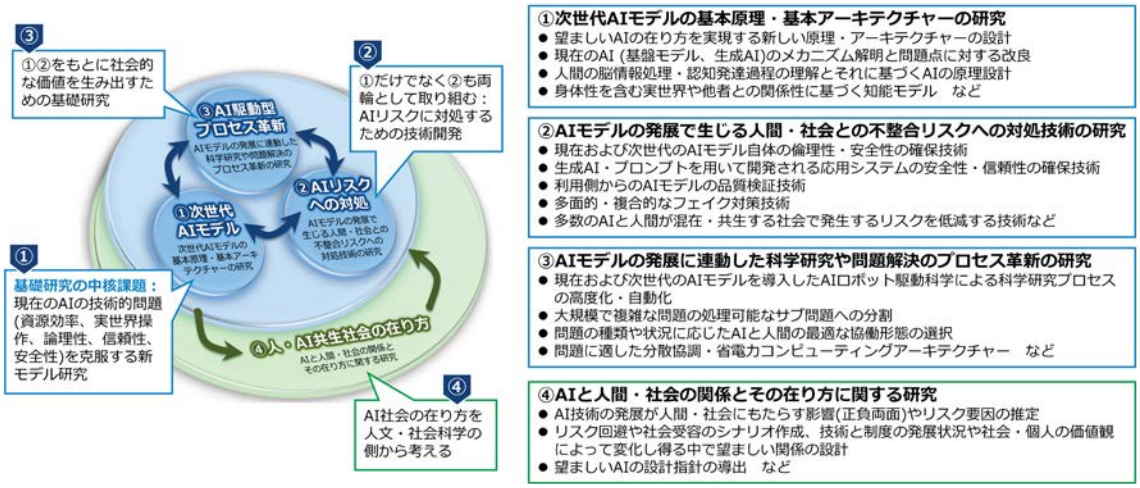


図3–1　研究開発課題の内容と課題間の関係（図1–1の再掲）

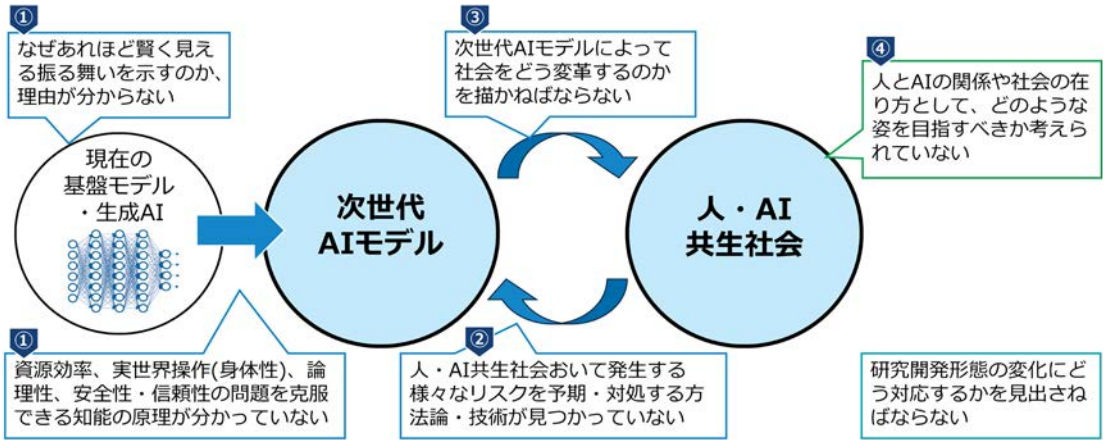


図3–2　解を見いだす必要のある問題と研究開発課題の対応

また、図3–2は、これらの課題への取り組みは、どのような問題に対して解を見いだそうとするものなのかという観点から、課題間の関係を捉え直したものである。まず、現在の基盤モデル・生成AIは、なぜあれほど賢く見える振る舞いを示すのか、理由が分からないので、それを解明するというのが、研究開発課題①の一つの側面である。また、現在の基盤モデル・生成AIは、資源効率、実世界操作（身体性）、論理性、安全性・信頼性に問題がある。それらの問題を克服する次世代AIモデルの新原理を見いだすというのが、研究開発課題①のもう一つの側面であり、特に中核的な課題である。一方、人とAIの関係や社会の在り方（人・AI共生社会）として、どのような姿を目指すかというのが、研究開発課題④の問いである。それを描きつつ、そこで発生するさまざまなリスクを予期・対処する方法論・技術を見いだそうとするのが研究開発課題②である。研究開発課題①②の両面から取り組み、生み出された次世代AIモデルによって、研究開発課題④で目指す社会へ変革するための取り組みが研究開発課題③である。なお、図3–2の右下に、研究開発形態の変化にどう対応するかを見いださねばならないという課題も載せたが、この課題を克服する施策については、第4章に示す。

以下、本章では、四つの研究開発課題の具体的な内容を説明する。

3.1 次世代AIモデル

研究開発課題①は、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である。より具体的な例示として、次のような取り組みが含まれる。

（1-1）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
（1-2）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
（1-3）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
（1-4）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

2.1.6（1）において説明したように、現在の基盤モデル・生成AIには、資源効率、実世界操作（身体性）、論理性、信頼性、安全性などの問題があり、次世代AIモデルでは、これらの問題の克服を目指す。そのためのシーズとなる技術開発として、以下のような取り組みが注目される。

（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
（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
（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
（d）知能を人間・社会との関係性の側面から発展させる

これらの取り組みの詳細は、2.1.6（1）および付録B.1で取り上げたが、2.1.6の表2–3に示したように、現状、それぞれは克服すべき問題に対して部分的な対処にとどまっている。

また、これらの取り組みの間で、目指すAIの姿に違いがあるようにも思える。図3–3に示すように、

（A）汎用性の高い道具としてのAI
（B）人間のパートナーとして望ましいAI
（C）人間の知能により近づいたAI

という方向性の差異がある。（A）は道具としての機能（例えば人間にはできないような機能を果たすなど）と制御性が重視されるのに対して、（B）はパートナーとして寄り添うような心理面が重視され、自律性も認められる。（A）と（B）は目指す姿は異なるものの、人間に役立つことを指向した工学的な立場であるのに対して、（C）は科学的な立場であり、人間の知能の理解・解明に動機がある。

現状、（A）（B）（C）の間には、方向性・立場の違いがあるが、徐々に融合・収束していくと見込まれる[18)]。その理由は以下の通りである。（A）の道具であっても、生成AIのように機能が高度化・汎用化すると、パー

トナーのような役割や自律性が高まり、（B）に近づく。（B）においても、自律性を認めるといっても、安全性・信頼性を確保するためには、一定の制約や制御性が求められる。また、（B）の設計においては、（C）の考え方や知見を取り込むことが有用である。

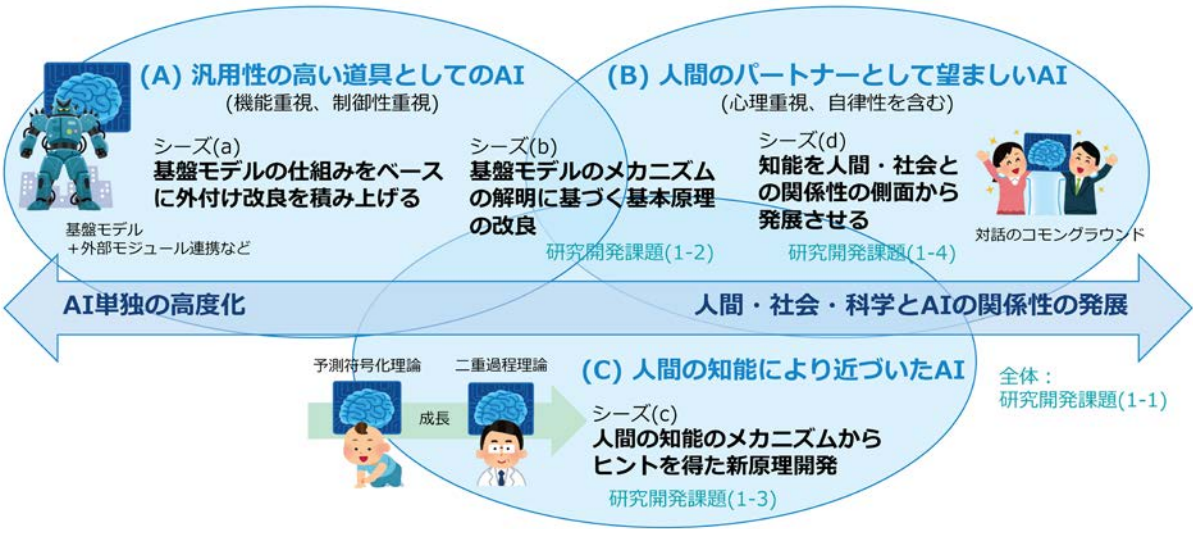


図3–3　次世代AIモデルに向けた幅広なアプローチ

全体としては、（A）から（B）へウェイトが移行しつつ、（C）を取り込んでいくことが見込まれる。このように（A）（B）（C）が徐々に融合・収束していくだろうこと、および、そこでシーズ技術となる（a）（b）（c）（d）の取り組みそれぞれでは、克服すべき問題への対処が限定的だが、それらを融合することで克服すべき問題に広く対処できる可能性があることから、**第1章**に示した方針A・方針Bの考え方の通り、次世代AIモデルに向けたアプローチの可能性を幅広に認めつつ、それらの取り組みの間で融合・シナジーが生まれるような進め方が有効と考えられる。また、幅広な可能性として、**図3–3**に示したように、AI単独での高度化だけでなく、人間・社会・科学との関係性の発展にも広げて考えることが重要になっていくと考えられる。

3.2 AIリスクへの対処

研究開発課題②は、AI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である。より具体的な例示として、次のような取り組みが含まれる。

（2-1）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
（2-2）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
（2-3）利用側からのAIモデルの品質検証技術
（2-4）多面的・複合的なフェイク対策技術
（2-5）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

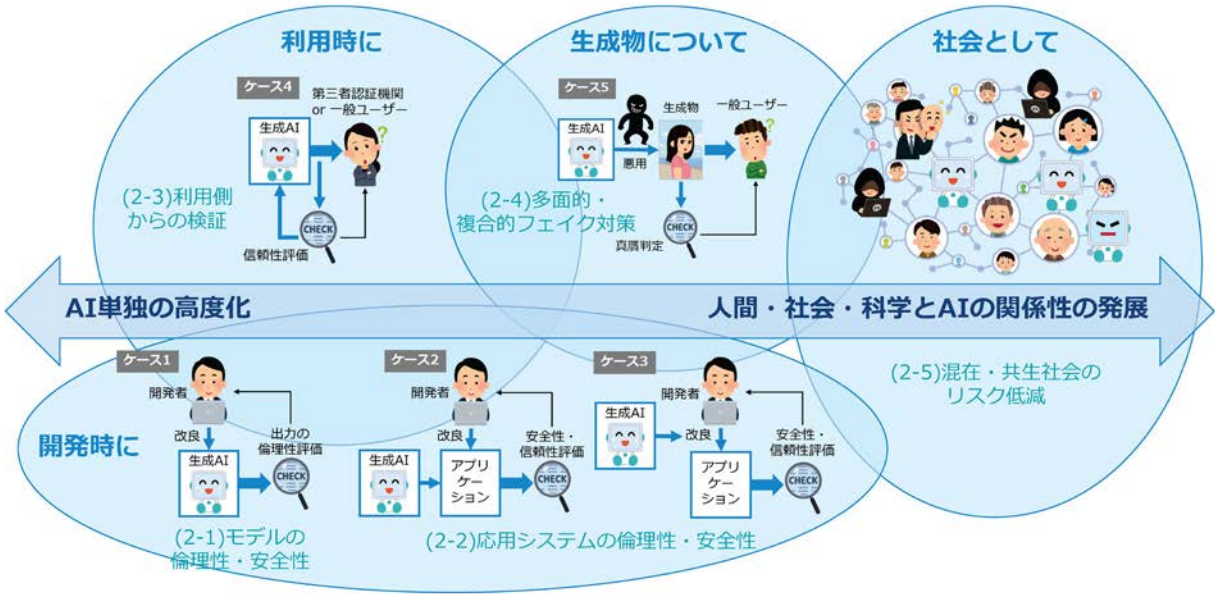


図3–4　　AIリスク対処場面の広がり

2.1.6（2）において説明したように、克服すべき問題・リスクは2.1.3の図2–5に示す通り多岐にわたる。2.1.6の図2–11では、それらへの対処パターンを五つのケースに分けて示したが、それらを図3–4では、AIモデルやAI応用システムの開発時、その利用時、および、生成物の流通・利用時という3種類にまとめて示した。

2.1.6で述べたように、現状、ケース1やケース5に対処する研究開発は活発化している一方、ケース2・3・4は問題が認識されつつあるものの、対処技術の研究開発はこれからである。また、ケース1やケース5は、喫緊の問題として研究開発が立ち上がったが、社会的な要因が絡み、判定条件に多様性や曖昧性があるため、本質的に難しい。したがって、これらいずれについても、継続して基礎研究を強化することが必要と考える。なお、関連する研究事例は付録B.1に挙げたことに加えて、ケース1の研究事例をコラム1に、新しい課題であるケース2・3に関わる問題の捉え方をコラム2に示した。

さらに、AIは精度が高まろうと原理的に100%の精度保証はできず、人間による誤用・悪用を完全に止めることも不可である。そのため、ケース1～5のような個別場面での技術的対処には限界がある。不安定なAIと不完全な人間が混在・共生する社会におけるAIリスク低減のための制度設計やインタラクション設計という、よりメタな問題に情報科学的なアプローチで取り組むことも必要と思われる。このような新しい研究課題に対して、マルチエージェント社会の問題として捉えたメカニズムデザイン手法やHuman-Agent Interaction設計論[22]、社会的トラスト形成[23]などの研究分野からのアプローチが考えられる。

このようなAIリスク対処場面の広がり（図3–4）を俯瞰すると、研究開発課題①において、AI単独での高度化だけでなく、人間・社会・科学との関係性の発展にも広げて考えることが重要になっていくという方向性が見られたが、この方向性が重要なのは研究開発課題②においても同様である。

コラム2

生成AIによるシステム開発のパラダイムシフト

システム開発のパラダイムシフトの歴史を図3–5のように捉えた。

第1パラダイムは、論理回路を組むことでシステムの動作を決める回路設計である。第2パラダイムは、手続きや判定ルールを明示的に書くことでシステムの動作を決めるプログラミングである。これら二つは演繹的な開発法になる。

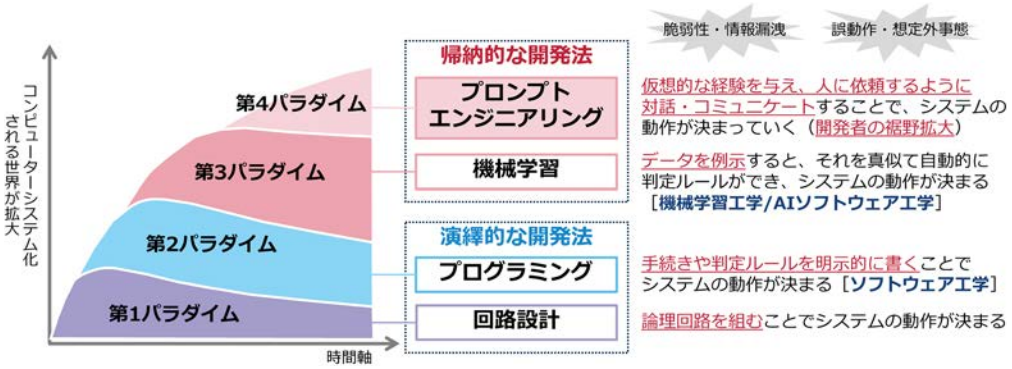


図3–5　システム開発のパラダイムシフトの歴史

それに対して、近年利用が拡大している機械学習を用いたシステム開発は、帰納的な開発法である。データを例示すると、それを真似て自動的に判定ルールができて、システムの動作が決まる。演繹的な開発法のために蓄積されてきた方法論や技術群がそのままでは活用できず、新たな方法論・技術群として、機械学習工学[24]（あるいはAIソフトウェア工学[25]）が整備されつつある。

さらに、生成AIを用いたプロンプトエンジニアリングによるシステム開発は新たなパラダイムと言えるかもしれない。生成AIは、ある意味、開発者の助手のような役割となり、生成AIに仮想的な経験を与え、人に依頼するように対話・コミュニケートすることで、システムの動作が決まっていく。自然言語で対話できるため、開発者の裾野が拡大することで、品質管理の方法論にも新たな側面が生じそうである。研究コミュニティーでも議論・検討が始まったばかりの新しい話題である。

3.3 AI駆動型プロセス革新

研究開発課題③は、AI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である。より具体的な例示として、次のような取り組みが含まれる。

（3-1）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
（3-2）大規模で複雑な問題の処理可能なサブ問題への分割
（3-3）問題の種類や状況に応じたAIと人間の最適な協働形態の選択
（3-4）問題に適した分散協調・省電力コンピューティングアーキテクチャー　など

2.1.6（3）において述べたように、AIロボット駆動科学（図2–12）に基盤モデルが適用されつつあるが、精度・効率をより高める上で、基盤モデルが抱える問題、特に論理性の問題と実世界操作（身体性）の問題の克服が望まれる。論理性の問題の克服・改善は、仮説生成・探索の精度を高めることにつながり、実世界操作（身体性）の問題の克服・改善は、検証実験に用いるロボットの柔軟で正確な制御を可能にする。

図3–6に示すように、現状のAIロボット駆動科学（Step 0）が科学研究サイクルの自動化を、ある程度限定された形式空間内で実現しているものと捉えると、これに基盤モデルを組み合わせること（Step 1）で、知識・言語空間との接続が可能になる。さらに、これに論理性と実世界操作が強化された次世代AIモデルを組み合わせること（Step 2）で、実世界や法則との接続が可能になると見込まれる。それによって、仮説空間探索の精度や効率が向上し、科学研究サイクルがより精緻で柔軟なものに革新される。

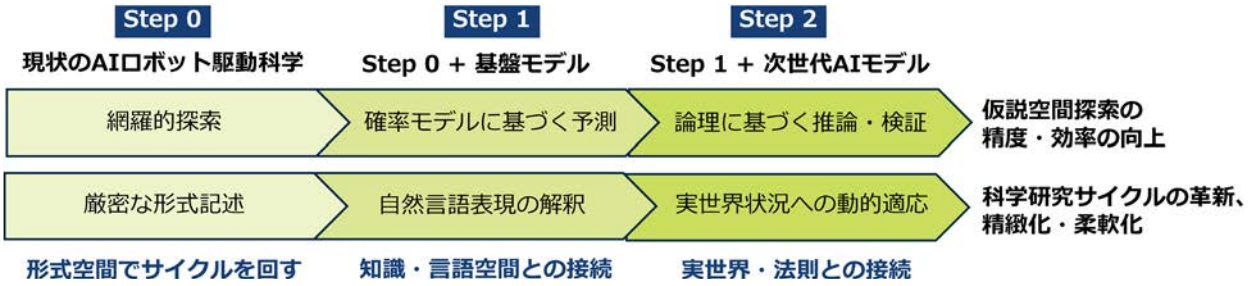


図3–6　基盤モデルと次世代AIモデルによるAIロボット駆動科学の発展

以上が、研究開発課題例の（3-1）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・自動化であるが、さらに、これまで述べてきたようなAIモデルの発展を、科学研究分野だけでなく、さまざまな分野の問題解決に広げて適用することを考えると、特に重要な研究開発課題例として、（3-2）大規模で複雑な問題の処理可能なサブ問題への分割、（3-3）問題の種類に応じたAIと人間の最適な協働形態の選択、（3-4）問題に適した分散協調・省電力コンピューティングアーキテクチャーなどが挙げられる。

現在の生成AIは、コンテキストに応じた予測応答を返すものであり、大きな枠組みとしてはパターン処理である。十分なコンテキスト情報が与えられれば、精度の高い予測応答を返すことができる。しかし、大規模で複雑な問題の場合、コンテキストとして関わる要素が非常に多く、また、ダイナミックに状況が変化することも多く、十分なコンテキスト情報を捉えにくい。そこで、大規模で複雑な問題のまま扱うのではなく、適切な規模のサブ問題に分割して処理する方策が有効と考えられる。そのために、どのような考え方・方針で問題を分割するのがよいか、研究開発課題例（3-2）が重要になる。

上記の分割されたサブ問題を含め、問題の種類に応じてAIの得意・不得意（精度の違い）があり、AIによる自動処理に委ねられるタイプのものもあれば、人による判断・アクションが不可欠なタイプのものもある。それを踏まえて、人とAIの役割分担や協働の仕方を適切に選択することが必要である。**コラム3**に示すように、人とAIの協働形態は大きく5タイプに整理できる。研究開発課題例（3-3）は、問題の種類や状況に応じてこれらを適切に選択するための技術開発である。

また、さまざまな問題にAI（現在のAIモデルおよび次世代AIモデル）の適用が広がっていくと、電力消費が大きな問題になる。さまざまな目的にカスタマイズされたAIが、多数乱立・共存する状況が予想される。次世代AIモデルでは省電力化を含む資源効率改善を可能にするようなアーキテクチャー刷新も期待されるが、市場全体を考えると、それで一気に解決するものではない。むしろ、AI処理（現在および次世代）に共通的に広く用いられる計算処理の省電力化や、さまざまな目的・場面に広く分散して実行されるエッジ側のAI処理の省電力化など、研究開発課題例（3-4）の問題に適した分散協調・省電力コンピューティングアーキテクチャーを考えていくことも重要である。

3.4 人・AI共生社会の在り方

研究開発課題④は、AIと人間・社会の関係とその在り方に関する研究である。より具体的な例示として、次のような取り組みが含まれる。

（4-1）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
（4-2）リスク回避や社会受容のシナリオ作成
（4-3）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
（4-4）望ましいAIの設計指針の導出　など

研究開発課題①②③は主に情報科学技術の研究であるのに対して、研究開発課題④は人・AI共生社会の在り方に関する人文・社会科学の研究である。**第1章**で述べたことだが、①②③④は以下のような関係になる。①が次世代AIモデル設計の中核になるが、②のAIリスクへの対処技術にも同時に取り組むことが不可欠である。①と②はいわば車の両輪のような関係となる基礎研究だが、そこから社会的な価値を生み出すための基礎研究が③のプロセス革新研究である。そして、**図3–1**にも示したように、④によって①②③の技術要件・指針が定まる一方、①②③の技術発展の結果として④の考え方や目標が変わり得る。

このような双方向性・スパイラル関係のもと、技術発展がもたらす社会の在り方への影響を考えるのが研究開発課題例（4-1）であり、社会の在り方から考えて望ましい技術発展の設計指針を与えるのが研究開発課題例（4-4）である。

（4-1）から技術発展がもたらす影響（正負両面）が考察されたとき、その影響をどう制御するかというシナリオ作りが重要であり、それが研究開発課題例（4-2）である。負の影響に対するリスク対策はその中心的なテーマの一つであるが、それだけでなく、時系列的に社会受容のシナリオを描くことも重要である。例えば、負の影響が少なく正の影響を享受しやすいようなユースケースを考えることができるならば、まずそのようなユースケースに限定して新技術を社会実装することで、社会受容が得られやすくなる。いったん正の影響（新技術の利便性など）が享受されると、人間・社会の側の価値観や受け止め方が変化し、その新技術に対する不安感が減少したり、リスク対策に新たな切り口が生まれたりし得る。

また、研究開発課題①②に見られたように、AI技術の発展は、AI単独での高度化だけでなく、AIと人間・社会・科学との関係性の発展という側面が強くなってくる。ここで望ましい関係とはどのようなものかは、種々の要因によって変化し得る。国や文化によって異なる社会の価値観に応じて、AIと人間・社会の望ましい関係は異なるものになるだろう。技術の発展と制度の発展は相互に影響し合うものだが、その状況によって、

AIと社会の望ましい関係は左右されるはずである。「AIが人間に寄り添う」というのは一見望ましいことのように思えるが、寄り添う相手がもし倫理的に問題のある価値観を持つ人間だった場合、AIはどのように振る舞うのが良いのかといった状況も考えられる。研究開発課題例（4-3）にはこのような多様な状況下での指針が期待される。

コラム3

人とAIの協働形態（Human-AI Teaming）

何らかの問題解決や目的達成に向けて人とAIが協働するときの組み方には、いろいろなバリエーションがあり得る。このような人とAIの協働について、国際規格ISO/IEC 22989：2022「Artificial intelligence concepts and terminology」では、Human-Machine Teamingという概念が示され、機械の知的な能力と人のインタラクションの統合（Integration of human interaction with machine intelligence capabilities）と定義されている[26)]。ここでいうMachineはAIのことであり、Human-AI Teamingとも呼ばれる[27)]。

文献[26)]では、人（Human）とAI（Machine）の関係を、それらの上下関係に応じて五つのタイプに整理している（図3–7）。人が上位となるタイプからAIが上位となるタイプの順に、Human Supervisor/User、Human Mentor、Peer、Machine Mentor、Machine Supervisorと名付けている。

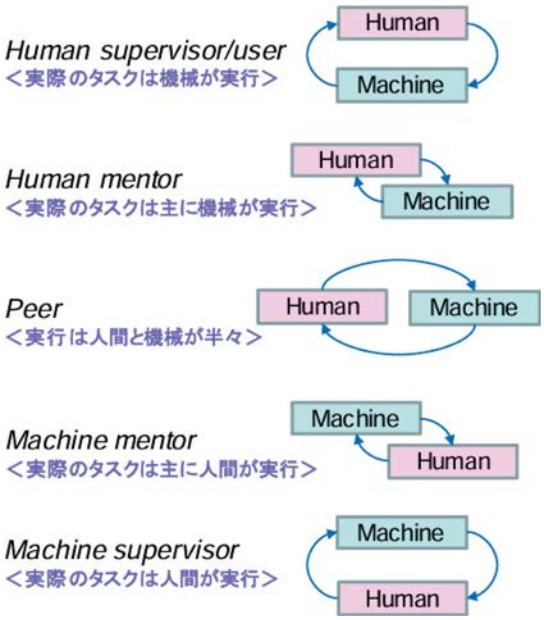


図3–7　Human-Machine（AI）Teamingの種類[26)]

一つ目の「Human Supervisor/User」タイプは、人がAIの上位に位置するパターンで、人が上司となるケース（Human Supervisor）と人が単なるユーザーのケース（Human User）がある。実際のタスクはAIによって実行される。医療画像診断などがHuman Supervisorのケースであり、人が責任を持って監督・介入するHuman Oversightが重要な課題である。その前提としてAIの透明性・説明性・制御性などが求められる。Human Userのケースでは、人がそこまで深く関与せず、一般にHuman Machine Interfaceが重要である。

二つ目の「Human Mentor」タイプは、人がAIのやや上位に位置するパターンで、AIが実際のタスクを主に実行し、人はそのタスクを実行しようと思えば実行できるものの、主としてMentorとして機械を指導する。問い合わせや検査などについて、AIで可能な範囲は自動処理して、難しいケースのみ人に対応させるというのが、その一例である。 AIへの権限移譲や人へのエスカレーションの仕方が重要課題である。そのために、人・AI双方が相手のモデルを持つこと（Mutual Model）が必要だと言われる。

三つ目の「Peer」タイプは、人とAIが同格で、どちらもタスクを実行する能力を有している。ただし、条件によってどちらのパフォーマンスが優れているかが変わってくるため、状況に応じてどちらがタスクを実行するかを決める必要がある。人とAIがタスクを分担して並列に実行することもあり得る。自動運転のレベル3はPeerに該当する。権限移譲の管理やMutual Modelが重要である。特に、人がAIの能力を適切に把握していることが望ましく、過信や不信を避けるように信頼較正[22)]という手法が考えられている。

四つ目の「Machine Mentor」タイプは、人が主としてタスクを実行するものの、一部のタスクに関してはAIが実行する。自動運転のレベル1や2はこれに該当する。 AIが人の作業・行動をモニタリングしていて、危ない状況や不適切な状況が検知されたら、注意や助言を行うケースもこのパターンの一例である。人のモチベーションへの配慮が求められる。

五つ目の「Machine Supervisor」タイプは、人が専ら実際のタスクを実行し、上司の立場にあるAIはタスクの実行には携わらない。ライドシェアサービスUberが代表例である。また、クラウドソーシングをAIで最適管理するようなケースも該当する。人のモチベーションを考慮したタスクアサインやフィードバックが課題として挙げられる。

なお、Human（人）、Machine（AI）ともに単独のケースも複数のケースも考えられる。また、実際の問題では、複数のタイプが組み合わせられることもある。各タイプにおいて、人とAIが協働する中で、人・AIそれぞれの能力が高まって、関係性・タイプが変化していくこともある。

コラム4

AIと哲学

AIの研究開発の目的・動機には大きく分けると二つの立場がある。一つは知能を解明したいという科学的な動機に基づくものである。これには、脳科学のように計測・分析によって脳の情報処理機構を明らかにしていこうというアプローチがある一方、実際にシステムを作ってみることで、知能に近づこうとする構成論的なアプローチもある。もう一つは、知的な処理を自動化することで、世の中に役立つものを提供したいという工学的な動機に基づくものである。空を飛ぶのに鳥と同様の仕組みを作るのではなく、飛行機を作るというような立場になる。

AI研究の黎明期には、科学的な動機から、知能とは何か、知能の働きをコンピューター上で実現できるのか、といった哲学的な議論が活発に行われたが、第3次AIブーム以降、AIを用いたさまざまなアプリケーションが実用化されるにつれて、工学的な動機からの取り組みが活発になっている。ただ、AI技術が急速に発展し、人間の能力に近づき、ある面では超えることが起こり始めたことで、AIと人間・社会との関係はどうあるべきか、（人間およびAIにおける）知能とは何か、といった哲学的な問いが再び沸き上がりつつある。

そのような動きを示す例として、人工知能学会が同学会誌『人工知能』の36巻1号（2021年1月）から38巻3号（2023年5月）にかけて、レクチャーシリーズ全11回[28]と総論記事3編[29]を掲載したことが挙げられる。レクチャーシリーズでは、毎回、AI研究者と哲学者が対談し、総論記事ではそれら対談を振り返りつつ、「哲学からAIへの15の批判」が示された。また、人工知能学会全国大会JSAI2023ではAI哲学マップ総括セッションが開催された。そのときの講演資料[30]が公開されているが、AI研究における哲学の意義、15の批判の内容、これまでの哲学の歴史・考え方とAI研究における概念・技術との対応などを含めて、詳細な検討がまとめられている。これらの企画を主導した三宅陽一郎の著書『人工知能のための哲学塾』シリーズ[31), 32), 33)]は、AIと哲学を扱った書籍としてよく知られている。

また、最近1年内に日本の哲学者から出版された書籍として、出口康夫の『AI親友論』[34]、鈴木貴之の『人工知能とどうつきあうか：哲学から考える』[35]、『人工知能の哲学入門』[36]がある。これらの間では、人間とAIの関係について、「主体としてのAI」と「道具（客体）としてのAI」という異なる立場を取っていることも興味深い。

本戦略プロポーザルの検討過程においても、研究開発課題④を中心に、哲学の視点を交えた議論が重要と考え、ワークショップとして「技術ブレークスルー編」だけでなく「AI×哲学編」を実施した。その詳細はワークショップ報告書[18]をまとめ公開するが、その中では「AI探求は人間探求、哲学によって深く探求し、エンジニアリングによって証明する」「AIの哲学による問題の再設定や思考枠の解体が役立つ」といった意見・期待が示された。また、各学問分野が細分化の傾向に

ある中で、哲学は諸学問に通じ、それらを広く俯瞰する視座を与え得ることにおいても、AIの発展に寄与するものと考える。

同報告書[18]においても言及しているが、哲学者や最先端AI研究者らの間で、AGI（Artificial General Intelligence）やASI（Artificial Super Intelligence）の可能性を含め、急速なAI技術の発展が招く人類滅亡のリスク（存在論的リスク）が議論されていることにも触れておきたい[37), 38)]。このような存在論的リスクへの対応を重視する立場は長期主義（後の世代の人が幸せに生きられるように、現代に生きる人たちは後世への影響を考えようという立場）につながる。このリスクに関して、欧米では以前から活発に論じられているのに対して、日本では議論が相対的に少ない状況だが、AI技術の急速な発展の中で重要な論点と考えておくべきであろう。危機的な状況に至るシナリオやその分岐条件も論じられているが[39)]、確実な回避策が見いだされているわけではない。汎用性の高い生成AIが登場したことによって問題意識が高まっており、欧米の長期主義を参照しつつも独自の検討を行う必要性の認識から、日本でも一般社団法人AIアライメントネットワークALIGN（AI Alignment Network）が2023年に発足した。

4 ｜ 研究開発の推進方法および時間軸

本章では、**第1章**で示した方針A～Dを踏まえて、**第3章**に示した研究開発課題①～④を推進するための方策を示す。

その際、特に方針Dで触れたような研究開発形態の変化を踏まえる必要がある。すなわち、AI分野の研究開発は、基礎研究においても、ビッグサイエンス化、ハイスピード化・ハイインパクト化、非オープン化の傾向が強まっている。基盤モデル・生成AIの研究開発において、わが国の取り組みが後追いになってしまったことについて、このような研究開発形態の変化に十分に対応できていなかったことが、大きな要因の一つとして考えられる。

研究開発形態の変化として挙げた3点は、以下のような傾向を意味する。

●ビッグサイエンス化

研究開発に必要な計算資源・データの超大規模化が進み、もはや大学研究室が単独で扱える規模ではなくなった。研究開発チームの規模も大型化し、チームを構成する人員は、研究者・科学者だけでなく、大規模計算資源や大規模データを高度に管理・強化するエンジニアや、大型チームのマネジメントや外部連携に関わる業務の支援スタッフなども含む。

●ハイスピード化・ハイインパクト化

数週間単位で注目技術が発表される。しかも、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生まれている。研究トレンドが短期で変化し、研究チーム組成の形態はよりアジャイルになり、かつ、研究の論文化に至るスパンがより短期化している。

●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。

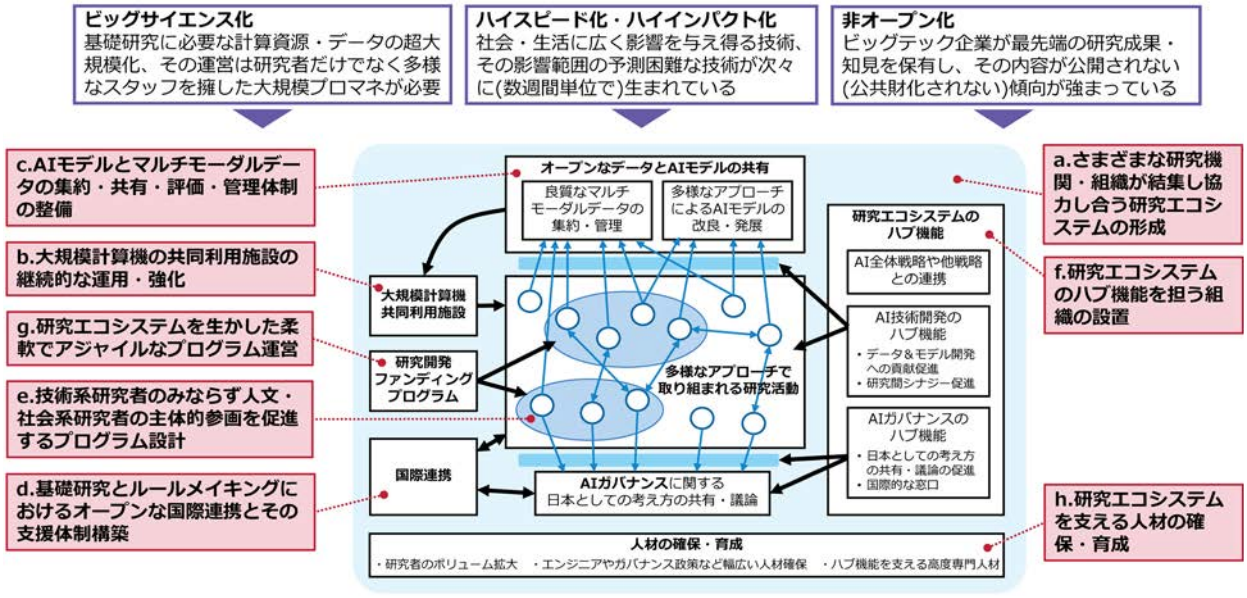


図4–1　研究エコシステムの概念と推進方法（図1–2の再掲）

このような状況下での研究開発体制・基盤やプログラムの在り方を考え、それを支える研究エコシステム（図4–1）をつくることが必要である。その中に含めるべき要素として、下記のようなものが挙げられる。補足として、ここで述べる研究エコシステムは、基礎研究の推進に主眼を置いたものである。産業・イノベーションのためのエコシステムに求められる要素と重なる部分は多々あるが、それを合わせて描くためには、より広いスコープからの検討が必要である。ただし、基礎研究の担い手としては、大学・国研などの研究機関だけでなく、企業の基礎研究部門も含めて考える。なお、下記の要素について、既に施策化が進んでいるものもあるが、研究エコシステムとして有機的につながる要素として全体像を示した。

a. さまざまな研究機関・組織が協力し合う研究エコシステムの形成

米国ビッグテック企業のような規模の研究開発体制や資金を確保することは、現状、日本の大学・企業では難しい。基礎研究において、多様なアプローチで取り組む研究機関が相互に協力し合う研究エコシステムを作っていくことが必要であろう。国研などがその中核機関になり得る。このような協力し合う体制を形成するには、インセンティブ設計が重要になる。特にcの項で述べるデータの共有については、自組織でコスト・労力をかけて集めたデータを共有することにためらいが生じがちであり、エコシステムへの貢献評価を含めたインセンティブをどのように設計するかが課題である。

b. 大規模計算機の共同利用施設の継続的な運用・強化

大規模計算機の共同利用施設は、研究エコシステムを構成する重要な要素である。最先端のAI基礎研究のための計算機環境は、大学などの一研究室で構築・運営できる規模ではなくなっている。日本では既に「ABCI」や「富岳」などが共同利用施設として稼働しているが、このような大規模計算機の共同利用施設の継続的な運用・強化が不可欠である。

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

研究エコシステムで開発したAIモデルをオープンな形で共有するとともに、その学習に用いるマルチモーダルデータの集約・共有も進め、それらを適切かつ効率よく管理する体制を構築・整備する。基盤モデルや次世代AIモデルが扱うデータは、自然言語や画像などの単一メディアデータからマルチメディア・マルチモーダルデータ、実世界データ、科学研究の多様な計測・実験データなどへと広がっており、今後の競争領域になっていくものと考えられる。その際、その種の学習データが必要になるが、単一メディアデータに比べて、それを集めることは容易でない。どのような環境・条件下でデータを集めるかを考えること自体が重要課題である。既に蓄積されている放送メディアや映像作品利用の可能性、オンラインへ移行が進んでいる各種業務の記録の仕方、ロボットなどを導入して自動化が進む工場や実験室でのデータ取得方法、ゲームやメタバースなどの仮想世界内の状況・行為のログデータの収集方法など、さまざまな切り口から検討し得る。

また、各組織で既に保有しているデータを共有・集約することで協力し合うだけでなく、それを解析して付与したアノテーションやメタデータも共有することで協力し合うデータエコシステム体制も望まれる。さらに、AIモデルの品質や学習データの品質を評価し改善していくための品質評価指標の確立や、その計測環境・ツールの整備も重要である。従来は設定したタスクの精度や達成率を評価指標とすることが多かったが、汎用性が高まったAIの評価はさまざまなタスクで評価する必要があり、どのようなタスク群を設定するのがよいかを考える必要がある。また、対話型生成AIの良さは、タスクの精度や達成率では捉えられない側面（利用する人間の心理面など）も関わってきそうである。

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

国内の大学・企業が連携した研究エコシステム体制を取りつつも、基礎研究やルールメイキングにお

いては、オープンに国際連携を推進することも重要である。現状、基礎研究においても、ルールメイキングにおいても、十分な人材が確保できていない。特にルールメイキングを含むAIガバナンス関連の国際連携においては、限られた人材に負荷が集中している。活動の国際的な窓口という意味では、少数の人材がその役割を担うのが関係継続に効果的だが、その場合でも、その少数人材を支援する体制の構築が必要になる。

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

既に述べてきているように、AI分野の研究開発には技術系研究者のみならず人文・社会系研究者の参画が不可欠である。その際、技術系研究者の手伝いや補助（例えば研究成果の最終フェーズでELSI面のチェックをするなど）ではなく、技術系の研究プロジェクトの初期フェーズから主体的に参画できることや、人文・社会系研究者の研究者が中核となる研究プロジェクトを確保することなど、人文・社会系研究者の主体的参画や連携機会の拡大を促進するプログラム設計も大事になる。また、技術系研究者として、情報技術分野が中心的に考えられがちだが、AI技術はもはや情報技術分野に限らず、あらゆる技術分野に関わるものであるから、データやモデルの構築・共有を含め、幅広い分野の技術系研究者の主体的な貢献が求められる。

f. 研究エコシステムのハブ機能を担う組織の設置

エコシステムを有効に稼働させるためには、ハブ機能が鍵となる。ハブ機能には複数の役割がある。まず、AI技術開発のハブ機能としては、データやAIモデルの開発にさまざまな研究機関・組織の貢献を促すことや、それらの研究活動の間のシナジーを促進することなどが役割となる。また、AIガバナンスのハブ機能としては、日本としての考え方の共有やそのための議論を促進することや、国際的な議論の場に対して窓口となることなどが役割となる。さらに、国のAI全体戦略や他の関連戦略・施策との連携を図ることも重要な役割となる。

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

AI分野は基礎研究においてもハイスピード化が進む中、最先端技術の状況に応じて、プロジェクトの目標を適宜見直すことも必要になるであろう。また、ファンディングプログラムの過度な縦割り運営が、共同利用施設やデータ集約・共有を進めるエコシステムの妨げにもなることも避けたい。柔軟でアジャイルなプログラム運営をどのように組み立てるかは課題である。

h. 研究エコシステムを支える人材の確保・育成

以上のような研究エコシステムに関わる取り組み全体を支える人材の確保・育成が不可欠である。まず研究人材のボリュームが足りていない。国内において情報系人材・AI人材を多く育成することが必要であるとともに、海外から優秀な人材が集まるような環境づくりも必要である。国際的な研究人材争奪戦の中で競争力のある処遇や国内外での研究開発経験を含めて描けるキャリア展望など、研究者視点から見た検討も課題であろう。また、研究エコシステムにおいては、研究者だけでなく、エンジニアやガバナンス政策なども含めた幅広い人材確保が必要である。また、ハブ機能を支える人材には、戦略策定や先端技術に深い理解を持つ高度専門人材の確保が必要である。このような多様な人材を確保・維持する上で、それぞれにとってのキャリアプランを設計することも重要である。

コラム5

AI研究のエコシステム形成に向けた施策の国内外事例

図4–1に示したようなAI研究のエコシステムを構築するため、各国はさまざまな政策手段を講じている。

英国においては、**アラン・チューリング研究所（The Alan Turing Institute）**[1]の存在感が大きい。2015年にデータサイエンスの国立研究機関として設立された同研究所は、国内の大学など30数機関と提携し、同国のAI研究のハブとなっている。2023年3月の新たな運営戦略「Turing 2.0」では、（1）世界レベルの研究推進とその社会課題[2]への応用、（2）データサイエンス・AI分野のスキル形成と人材育成、（3）技術・社会・倫理的側面の市民対話の促進、を同研究所の三つのゴールとして打ち出している。

米国においては、**国立AI研究所（National Artificial Intelligence Research Institute）**のネットワーク形成が進む。米国政府は2020年頃から国立科学財団（NSF）などを通して累計5億ドル近くを投じ、大学を中心とした25の国立AI研究所を設立してきた。各拠点では、「サイバーセキュリティーのための人工エージェント」「気候変動に対応した林業と農業のスマート化」「意思決定のためのAI」といった特色のあるトピックに取り組む。加えて、2023年末から開始したNAIRR（国家AI研究リソース）プログラム[3]を通じ、全米のアカデミア機関が活用できる研究資源（データや計算資源）の整備を進めようとしている。

英国と米国はこのように異なるアプローチで公的なAI研究拠点のネットワーク形成を進めているが、どちらもAIの技術開発のみならずAIガバナンス（付録B.2参照）につながる、人文・社会科学も含む広範な分野の研究・実践も包含している。アラン・チューリング研究所は、AIをめぐる標準化や国際的なルール形成にも強い影響力を持つ[4]。また、米国の国立AI研究所の一つであるTRAILS Institute（中心機関：メリーランド大学）は、人文・社会科学を核に、参加型のAI設計、AIの信頼性評価、包括的なAIガバナンス研究などを進めている。また、両国ともに、長期的な基礎研究の促進、イノベーションの実現に向けた学際的・多層的な体制などの構築、次世代人材の育成などが一体的にデザインされている点に特徴がある。

日本国内にも、AI研究の中核機関が存在している。理化学研究所の革新知能統合研究センター（AIPセンター、2015年設立）、産業総合研究所の人工知能研究

1 アラン・チューリング研究所に所属の研究者は400人超（多くは連携機関との兼務）で、年間予算は£52.5m（～95億円、うち£30.8mが公的資金）[41)]。

2 健康、サステナビリティ、防衛・国家安全保障を三つの「グランドチャレンジ」としてAI応用の重点分野としている。

3 NAIRRは11の政府機関と25の民間パートナー機関により、2024年1月にパイロットプログラムを開始した。 The National Artificial Intelligence Research Resource（NAIRR）Pilot https://nairrpilot.org/

4 例えば同研究所のEthics and Responsible Innovation Researchチームによる論文が、欧州評議会の人工知能に関する特別委員会（CAHAI）でのAI条約の議論でのたたき台となるなど、条約交渉で中心的な役割を担っている。

センター（AIRC、2015年設立）、情報通信研究機構（NICT）といった国研の他、2024年には国立情報学研究所（NII）に大規模言語モデル研究開発センター（仮称）が設立される見込みである。日本においても、これらの中核機関と、大学などの拠点、ならびにAI分野の研究開発事業・プログラムとの連携が進むことが望ましい。そうした相互連携機能を実現すべく実施されている取り組みとして下記がある。

- **AIPネットワークラボ：** 2016年度から文部科学省が開始したAIPプロジェクトの実施機関として、理化学研究所AIPセンターとともに、JST戦略的創造研究推進事業の関連する研究領域群で編成し、統合的な研究拠点と独創的な研究を支援・加速するファンディング事業とを一体的に推進する枠組み。これまで、フランスの資金提供機関との共同公募などの国際連携、若手人材育成事業、ネットワークラボ所属領域のプロジェクト終了後の加速研究の支援など、JST戦略事業がファンドするICT関連のプロジェクトを横断する事業を展開してきた。
- **人工知能研究開発ネットワーク（AI Japan R&D Network）：** 2019年12月に産総研、理研、情報通信研究機構（NICT）を中核会員として設立されたコンソーシアムをもとに、2023年4月に民間事業なども含めた任意団体に改組された連携の枠組み。機関ごとに窓口をつくってネットワーク化している。2020年6月には、このネットワークを活用して「AIを使ったコロナ対策」に関する研究を70件リストアップし、内閣府にて活用されるなどの実績を持つ[5]。

上記のような日本の既存の連携機能やネットワークをリソースとしつつ、ビッグサイエンス化、ハイスピード化・ハイインパクト化、非オープン化という潮流に対応すべく前述のa～hの観点に配慮した国内のAI研究体制をアジャイルに組み立て、動かしていく必要がある。

4 研究開発の推進方法および時間軸

推進方法の時間軸については、技術開発のスピードが速いことと、その状況変化の中で複数の施策が並行して検討・実施されていることから、上に挙げたような施策は、全体のロードマップを描いて取り組むというよりは、状況に合わせてアジャイルに推進する形が適する。 AI分野の研究開発形態の変化への対応は急がねばならないが、上に挙げたような要素全てを詳細に設計して一気に実行に移すことは難しい。既に動きつつあるものもあり、全体ビジョンを共有しつつ、ボトムアップに組み上げていくのが現実的である。また、基礎研究のエコシステム構築は、国の政策が主導するというより、大学などの研究機関や企業の基礎研究部門の主体的な参画をベースに考えるのが良い。国の政策としては、エコシステムの中核機関になる組織の活動や人材育成・確保に対して支援することが期待される。

また、本プロポーザルでは、2.1.5で述べた通り、基礎研究の戦略にフォーカスした。現在の基盤モデルベースのAIの限界・問題を克服する次世代AIモデルの基本原理・アーキテクチャーがまだ必ずしも定まっていない状況で、現状の技術シーズを足掛かりに幅広に可能性を探索・探求するフェーズでの戦略を中心に示した。

5 人工知能研究開発ネットワーク「新型コロナウイルス感染症対策に係るAIを活用した取組」
https://www.ai-japan.go.jp/menu/covid-19-top/covid-top/

この幅広な可能性の探索・探求フェーズを通して、特に有望な基本原理・アーキテクチャーが見いだされたならば、それに重点化するように戦略を切り替えることが考えられる。

ただし、そのような基礎研究の成果だけで、産業・経済の大きな成長やイノベーションが起こせるわけではない。しかし、産業・経済の成長やイノベーションを加速する要因になり得る。基礎研究の戦略・推進策とイノベーションのための戦略・推進策は、車の両輪のように連動させて進めていく必要がある[6]。図4–2は、2.1.5で示した課題の全体像の図（図 2–7）に、本プロポーザルで示した基礎研究の戦略（2）と、それ以外の取り組み（1）（3）との関係、および、それらを連動して進めることの必要性を示したものである。

図4–2の上部（1）は、基盤モデル・生成AI活用によって生産性向上DX・事業成長を促進する取り組みである。これについては、産業界主導で活発に取り組まれているとともに、図 2–6に示した国の政策においても「AIの利活用促進」が掲げられている。右部（2）は本プロポーザルが対象とした、次の世代のAIモデルで先行を狙う基礎研究の推進である。大学・アカデミアおよび企業の基礎研究部門がけん引する。それを国のファンドによって加速する。これら（1）と（2）は上述したように両輪の関係であり、（1）の収益が（2）への投資になるとともに、（2）の成果が（1）を加速・拡大する。

図4–2の左下部（3）は、（1）や（2）を支えるインフラ・実行環境となるが、現状は米国ビッグテック企業の後追いの状況で、追い抜くことは難しい。しかし、（1）（2）を支え連動しつつ、徐々に底上げされていくと期待される。この部分は、ビジネス用と研究用があり、ビジネス用については、基本は産業界主導で取り組むものだが、必要に応じて国が促進策を講じることもあり得る。研究用については、国の支援による共同利用施設の整備が求められる。

以上のような、より上位の施策設計との連動を考えて、本プロポーザルで示した次世代AIモデルに向けた基礎研究の戦略を推進していく必要がある。

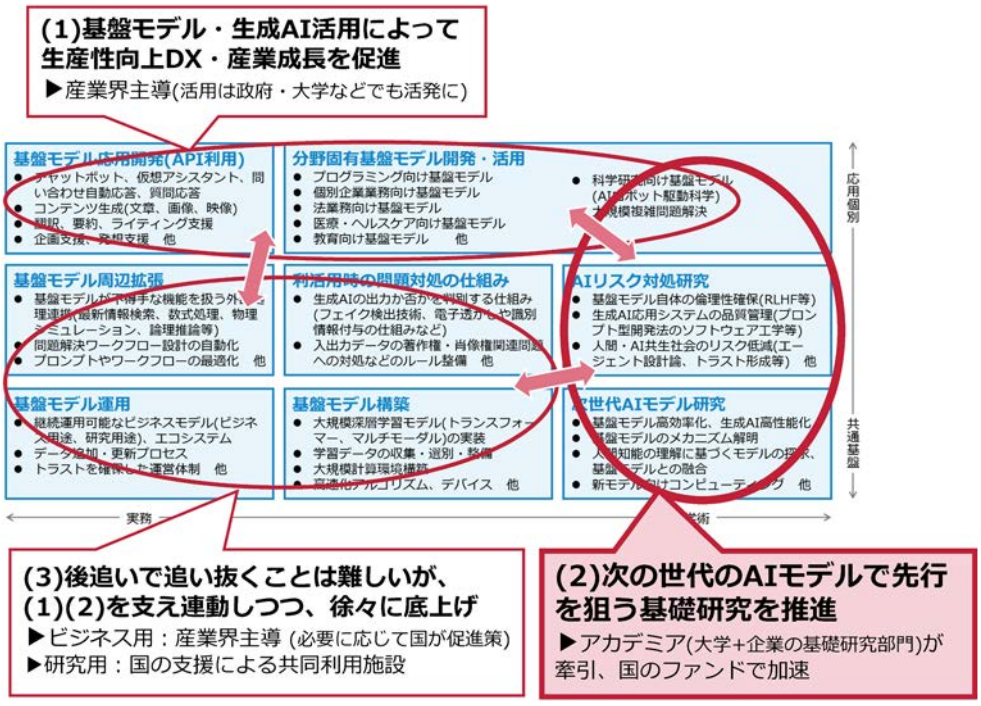


図 4–2　　基礎研究以外の取り組みとの連動

6　図2–6はこれに関連する国の戦略を示したものである。また、本プロポーザルの検討に関わる科学技術未来戦略ワークショップ[18]においても、イノベーションのための戦略・推進策や基礎研究の戦略・推進策との関係についての議論や提案が示された。

付録

付録A 検討経緯

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）では、2023年度の戦略スコープ策定委員会において、戦略プロポーザルを作成すべきテーマの候補として、本テーマを選定し、検討チームを発足させた。本検討チームは2023年7月から活動を開始し、2024年2月まで検討を進めてきた。

その間、専門家へのインタビューや、ワークショップを実施し、本テーマの研究開発状況の把握や研究開発課題・方向性の議論を進めてきた。また、非常に動きの速いテーマであるため、チームでの検討と並行して、政策関係者や研究コミュニティーとの意見交換を進めてきた。以下に、その概要を示す。

A.1 科学技術未来戦略ワークショップ （詳細は報告書[18] 参照）

日程：第1回［技術ブレークスルー編］2023年11月23日（木）13：00～17：00
第2回［AI×哲学編］ 2023年12月20日（水） 9：00～12：30

場所：Zoom Meetingによるオンライン開催

第1回［技術ブレークスルー編］ワークショップのプログラム：（敬称略）

［第1部］13：00～13：20

・開催挨拶 木村康則（CRDS）
・提言骨子の紹介と論点の説明 福島俊一（CRDS）
【論点1】現在の基盤モデル・生成AIの本質的な問題点は何か？
【論点2】その問題点を解決する次世代AIモデルへの重要なアプローチは何か？
その世界的な動向や日本のポジションは？
【論点3】米中2強で劣勢の日本が取り得る戦略・方策は？

［第2部］13：20～15：50 論点および関連動向・取り組みに関する話題提供（各発表15分＋質疑）

・13：20～13：45 黒橋禎夫（国立情報学研究所）：LLM-jp：自然言語処理およびLLMメカニズム理解への取り組み
・13：45～14：10 東中竜一郎（名古屋大学）：対話システムにおけるブレークスルー
・14：10～14：35 牛久祥孝（オムロンサイニックエックス株式会社）：画像処理とAIロボット駆動科学
・14：35～15：00 山川宏（全脳アーキテクチャ・イニシアティブ）：AIと人類のアライメントを橋渡しするヒト脳型AGI
・15：00～15：25 尾形哲也（早稲田大学）：認知発達ロボティクス、身体性・実世界相互作用を踏まえた知能
・15：25～15：50 松尾豊（東京大学）：松尾研究室の活動紹介と生成AIについて

［第3部］（15：50～16：00休憩の後）16：00～17：00

・総合討論 司会：福島俊一（CRDS）
・ディスカッサント：大森久美子、高橋恒一、谷口忠大、銅谷賢治、村上祐子

第2回［AI×哲学編］ワークショップのプログラム：（敬称略）

［第1部］9：00～9：15

・開催挨拶　　木村康則（CRDS）
・提言骨子の紹介と論点の説明　　福島俊一（CRDS）
【論点1】人工知能学会「AI哲学マップ」[30] の「哲学からAIへの15の批判」を起点として、以下の1a・1bを深掘り（15の批判に関する異論・追加・具体化や生成AI登場による変化なども）
【論点1a】現在のAIの知能としての根本的課題、次世代AIモデルに求める要件
【論点1b】AIと人間・社会の相互作用（スパイラル）、AIと人間・社会との関係のあるべき姿
【論点2】このような議論を継続し、AI研究に反映していくための学際的研究の進め方・課題
［第2部］9：15～11：50　論点および関連動向・取り組みに関する話題提供（各発表15分）
・ 9：15～ 9：30　三宅陽一郎（株式会社スクウェア・エニックス）：AI哲学マップ
・ 9：30～ 9：45　鈴木貴之（東京大学）：人工知能の哲学2.0
・ 9：45～10：00　谷口忠大（立命館大学）：記号創発システム、集合的予測符号化
・10：00～10：15　田口茂（北海道大学）：「人工主体」の生成？　身体性と生命性・死の意義
・10：15～10：45　前半の話題提供に対する質疑＋休憩
・10：45～11：00　大森久美子（京都哲学研究所/NTT）・高木俊一（京都哲学研究所/京都大学）：京都哲学研究所のビジョン
・11：00～11：15　村上祐子（立教大学）：哲学とAI
・11：15～11：30　高橋恒一（理化学研究所）：先端AIと長期主義に関する批判的考察とAIアライメント
・11：30～11：50　後半の話題提供に対する質疑
［第3部］11：50～12：30
・総合討論　　司会：福島俊一（CRDS）
・ディスカッサント：牛久祥孝、尾形哲也、黒橋禎夫、銅谷賢治、山川宏

聴講参加者（上記参加者以外）：

第1回	内閣府	4名	
	文部科学省	5名	
	経済産業省	1名	
	NEDO	7名	
	JST CRDS	16名	
	JST CRDS以外	5名	合計　38名
第2回	内閣府	4名	
	文部科学省	8名	
	経済産業省	1名	
	NEDO	6名	
	JST CRDS	19名	
	JST CRDS以外	14名	合計　52名

A.2　専門家インタビュー

相澤 彰子	国立情報学研究所 コンテンツ科学研究系 教授/所長代行/副所長
合原 一幸	東京大学 特別教授/ニューロインテリジェンス国際研究機構 副機構長
青木 孝文	東北大学大学院情報科学研究科 教授/理事/副学長
浅田 稔	大阪国際工科専門職大学 副学長
荒木 正太	一般社団法人京都哲学研究所/日本電信電話株式会社

五十嵐 涼介	京都大学大学院文学研究科 特定講師/一般社団法人京都哲学研究所 研究員
石川 冬樹	国立情報学研究所アーキテクチャ科学研究系 准教授
伊藤 孝行	京都大学大学院情報学研究科 教授
乾 健太郎	Mohamed bin Zayed University of Artificial Intelligence（MBZUAI）, NLP Department, Visiting Professor/東北大学大学院情報科学研究科 教授
植田 一博	東京大学大学院総合文化研究科 教授
牛久 祥孝	オムロンサイニックエックス・プリンシパルインベスティゲーター
越前 功	国立情報学研究所 情報社会相関研究系 教授/主幹
江間 有沙	東京大学国際高等研究所東京カレッジ 准教授
大塚 淳	京都大学大学院文学研究科哲学専修 准教授
大森 久美子	一般社団法人京都哲学研究所 理事/日本電信電話株式会社
岡崎 直観	東京工業大学情報理工学院 教授
尾形 哲也	早稲田大学理工学術院基幹理工学部 教授
小山田 昌史	日本電気株式会社 NECデータサイエンスラボラトリー 主幹研究員/ディレクター
小川 秀人	日立製作所研究開発グループ システムイノベーションセンタ 主管研究長
神嶌 敏弘	産業技術総合研究所 情報・人間工学領域 主任研究員
神谷 之康	京都大学情報学研究科 教授/ATR情報研究所 客員室長（ATRフェロー）
川名 晋史	東京工業大学リベラルアーツ研究教育院 教授
川原 圭博	東京大学大学院工学系研究科 教授/インクルーシブ工学連携研究機構 機構長
久木田 水生	名古屋大学大学院情報学研究科社会情報学専攻 准教授
工藤 郁子	大阪大学社会技術共創研究センター 特任研究員
栗原 聡	慶應義塾大学理工学部 教授
黒橋 禎夫	国立情報学研究所 所長
國領 二郎	慶應義塾大学総合政策学部 教授
小宮山 純平	NYU Stern School of Business, Assistant Professor
佐倉 統	東京大学大学院情報学環 教授
杉山 将	理化学研究所 革新知能統合研究センター
鈴木 貴之	東京大学大学院総合文化研究科 教授
鈴木 雅大	東京大学大学院工学系研究科 松尾研究室 特任助教
須藤 修	中央大学国際情報学部 教授/中央大学ELSIセンター 所長
平 和博	桜美林大学リベラルアーツ学群 教授
高橋 恒一	理化学研究所バイオコンピューティング研究チーム チームリーダー
田口 茂	北海道大学 大学院文学研究院 教授/人間知・脳・AI研究教育センター センター長
谷口 忠大	立命館大学情報理工学部 教授
千葉 雄樹	日本電気株式会社 NEC Generative AI Chief Navigator
辻井 潤一	産業技術総合研究所 フェロー
銅谷 賢治	沖縄科学技術大学院大学神経計算ユニット 教授
中川 裕志	理化学研究所革新知能統合研究センター チームリーダー
中島 秀之	札幌市立大学 学長
原田 香奈子	東京大学 大学院医学系研究科/大学院工学系研究科 准教授
原田 達也	東京大学先端科学技術研究センター 教授
東中 竜一郎	名古屋大学大学院情報学研究科 教授
牧野 貴樹	Google Inc. Staff Software Engineer

丸山 文宏	産業技術総合研究所社会実装本部 招聘研究員
三宅 陽一郎	株式会社スクウェア・エニックス リードAIリサーチャー
村上 祐子	立教大学大学院人工知能科学研究科 教授
山川 宏	全脳アーキテクチャ・イニシアティブ 代表
山本 貴光	東京工業大学リベラルアーツ研究教育院 教授
Sam Passaglia	東京大学カブリ数物連携宇宙研究機構

A.3　政策関係者や研究コミュニティーとの意見交換

基盤モデル・生成AIまわりの研究開発および政策検討の動きが速いことを踏まえて、2023年7月にスタートしたチーム活動に先立ち、本テーマに関連して既に調査していた分野動向や方向性の示唆について報告書「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～」[7)]を2023年7月に公表した[1]。この報告書の内容に基づく講演依頼や関連トピックの問い合わせ・取材なども多く、そこでの意見交換・フィードバックも本プロポーザルの検討に取り込ませていただいた。府省の委員会などでの発表もあり、先行して政策検討にも活用いただいている。

政策関係者へのプレゼンテーション

- 2023年5月10日に開催された文部科学省内のオンライン研修にて講演
- 2023年6月21日に開催された文部科学省科学技術・学術審議会第11回基礎研究振興部会[2]にて発表
- 2023年6月23日に開催された科学技術と経済の会（FF会）[3]にて講演
- 2023年7月4日に文部科学省初等中等教育局から公開された「初等中等教育段階における生成AIの利用に関する暫定的なガイドライン」[4]の作成に協力
- 2023年11月22日に開催された第236回科学技術政策懇談会にて講演
- 2024年1月25日に開催された科学技術・学術審議会第35回情報委員会[5]にて発表

研究コミュニティーへのプレゼンテーション

- 2023年7月12日に開催された理化学研究所革新知能統合研究センター「AIと文化シンポジウム：生成系AIの活用」[6]にて招待講演
- 2023年10月31日に開催されたコンピュータセキュリティシンポジウムCSS 2023のAIセキュリティー企画セッション[7]にて招待講演
- 2023年11月25日に開催されたJST CREST「信頼されるAIシステム」[8]領域会議にて招待講演
- 2024年1月28日に開催された栢森情報科学振興財団AIシンポジウム「AI：夢が現実に、夢を未来に ～AI新世紀を語る～」[9]にて招待講演（午前：クローズド）とパネル討論（午後：公開）

1 この報告書[7)]は、2021年6月に公表した報告書「人工知能研究の新潮流 ～日本の勝ち筋～」[40)]を、基盤モデル・生成AIを含む最新動向を盛り込んでアップデートしたものである。
2 https://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu27/index.htm
3 https://www.jates.or.jp/management_study/ff.html
4 https://www.mext.go.jp/content/20230704-mxt_shuukyo02-000003278_003.pdf
5 https://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu29/index.htm
6 https://aip.riken.jp/events/event_157845/
7 https://www.iwsec.org/css/2023/program.html#2B1
8 https://www.jst.go.jp/kisoken/crest/research_area/ongoing/bunya2020-4.html
9 https://kzaidan.org/

なお、上記のような講演のための参加以外に、動向調査のための国内外の学会参加や講演会・セミナー聴講（100件程度）と、その中での専門家との意見交換も重ねてきた。

付録B 国内外の状況

B.1 研究開発課題に関わる技術開発事例

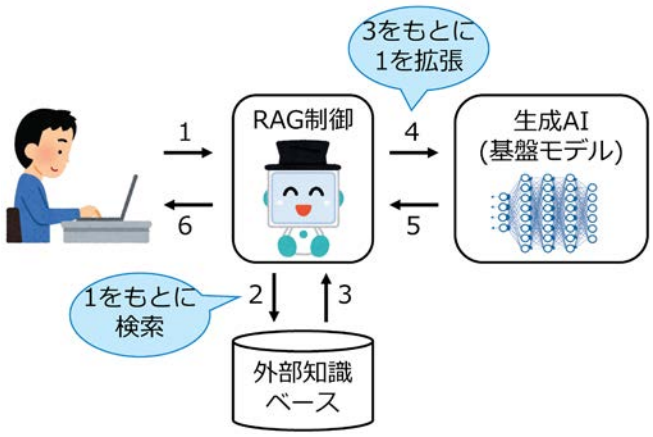
本節では、研究開発課題①「次世代AIモデル」と研究開発課題②「AIリスクへの対処」に関連した技術開発事例から代表的なものを挙げる。

B.1.1 次世代AIモデルの技術シーズ

（1）現在の基盤モデルの仕組みをベースに外付け拡張

現在の基盤モデルは、事前学習されたトランスフォーマー型の深層ニューラルネットを確率モデルとして用いて、続きを予測するものであるために限界があり、その外側に機能拡張・改良の仕組みを付けることが行われている。その仕組みの代表的なものとして、プラグインやLangChainなどが知られており、既にこれらを用いたアプリケーションが多数開発されている。例えば、現在の基盤モデルは数式計算が苦手なことが知られているが、それを数式処理用の外部モジュール（Wolframプラグインなど）に受け持たせるといったことが行われている。 LangChainは外部リソースと連携させる仕組みで、例えば、基盤モデルは事前学習した時期までの情報しか持っていないので、欠落している最新情報をサーチエンジンと連携して補うようなことが可能になる。

このような仕組みなどを用いつつ、生成AIをうまく使う方法の例として、AutoGPTやRAGなどが知られている。 AutoGPTは、検索やファイル入出力なども含むワークフローを、目的タスクに合わせて自動設計して実行してくれる。例えば、あるトピックについて調べて解説記事をまとめるという目的タスクであれば、そのトピックについてサーチエンジンを検索した結果をいったんファイルに出力し、その内容を生成AIに読み込ませて要約を別ファイルに出力させるといった一連のワークフローを自動設計し、必要であれば、途中ステップでユーザーに確認を求めながら実行する。 RAG[42] は、Retrieval Augmented Generation（検索拡張生成）の略で、ユーザーからの入力（プロンプト）を生成AIに渡す前に別途用意した外部知識ベース（例えば専門知識や社内情報など）を検索し、プロンプトだけでなく検索結果から抽出した情報も合わせて生成AIに渡して応答を生成するようにするというものである（図B–1–1）。目的タスクに応じて、よりカスタマイズされた応答が可能になり、ハルシネーションの抑制効果も見込まれている。



図B–1–1 RAGの処理フロー

このような外付け拡張のアプローチは、導入が容易であり、さまざまなビジネス応用を含めて活用が広がっている。ただし、現在の基盤モデルの問題の一つとして挙げた資源効率の問題に対しては、根本的な解決にはならず、より多くの計算資源が必要になり、むしろ問題が悪化するように思われる。

（2）基盤モデルのメカニズム解明の取り組み

基盤モデルは、確率モデルに基づいて予測する仕組みでありながら、あれほど賢いように見える振る舞いをするのか、そのメカニズムは明らかになっていない。基盤モデル（大規模言語モデルLLMを含む）のメカニズムを解明することは、現在の基盤モデルの問題の原因を明らかにし、モデルを改良・発展させるための基礎となる重要な取り組みである。これまでも深層学習の原理解明のための研究が進められてきたが[43]、より大規模なモデルにおける挙動や現象の解析・メカニズム解明への取り組みが必要である。

その代表的な活動として、国立情報学研究所（所長：黒橋禎夫）を中心として、大学・企業などから700名以上が参加しているLLM-jp（LLM勉強会）が挙げられる。オープンかつ日本語に強い大規模モデルを構築し、LLMの原理解明に取り組んでいる。モデル、データ、ツール、技術資料などを、議論の過程や失敗も含めて全て公開する方針で進めている。コーパス構築WG、モデル構築WG、評価・チューニングWG、安全性WGなどのワーキンググループ（WG）活動も行われており、2023年10月に130億パラメーターの大規模言語モデルLLM-jp-13Bを公開した。さらに、2023年9月に産業技術総合研究所ABCIの第2回大規模言語モデル構築支援プログラム、2024年2月に経済産業省のGENIACプロジェクト（2.1.2にて言及）にも採択され、GPT-3（1750億パラメーター）級の大規模言語モデルの構築を進めている。2024年4月には国立情報学研究所に大規模言語モデル研究開発センター（仮称）の設立も予定されている。

（3）二重過程理論に基づくAIモデル設計への取り組み

人間の脳には、深層学習・強化学習に相当するような、経験に基づいた即応的な情報処理を担う部分だけでなく、ある種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部分が存在すると考えられている。行動経済学で有名なDaniel Kahnemanは、著書「ファスト＆スロー」[44]において、人間の知能（意思決定システム）には二つの側面があるとし、反応の早い即応的知能をシステム1、それに比べると反応が遅い熟考的知能をシステム2と呼んだ（2.1.6の表2–4）。それ以前から社会心理学や認知心理学などの心理学分野では「二重過程理論」（Dual Process Theory）と呼ばれていた。

機械学習分野のトップランク国際会議として知られるNeurIPS（Conference on Neural Information Processing Systems：神経情報処理システム会議）の2019年開催における基調講演の一つは、カナダのモントリオール大学教授Yoshua Bengioによる「From System 1 Deep Learning to System 2 Deep Learning」[10]と題したものであった。現在の深層学習はシステム1に相当し、将来の方向性はシステム2の深層学習だというのが、この講演のメッセージである。

従来の深層学習はシステム1で実行される処理に相当すると考えられたが、基盤モデルはシステム2も部分的にカバーしているように思われる。メカニズムの詳細は明らかになっておらず、論理構築・論理推論が苦手なことから、システム2の実現には至っていない（図B–1–2）。帰納型の処理だけで精度を高めようとすると、どうしても大量データからのボトムアップ学習が必要になり、資源効率の問題は回避できない。演繹型の推論を用いるならば、必要なデータのみ取りに行くというトップダウン制御が可能になり、資源効率の問題に対処できようになる。また、論理性の問題にも対処できるとともに、トップダウンの制御によって安全性の問題にも対処しやすくなるはずである。

10 https://www.youtube.com/watch?v=T3sxeTgT4qc

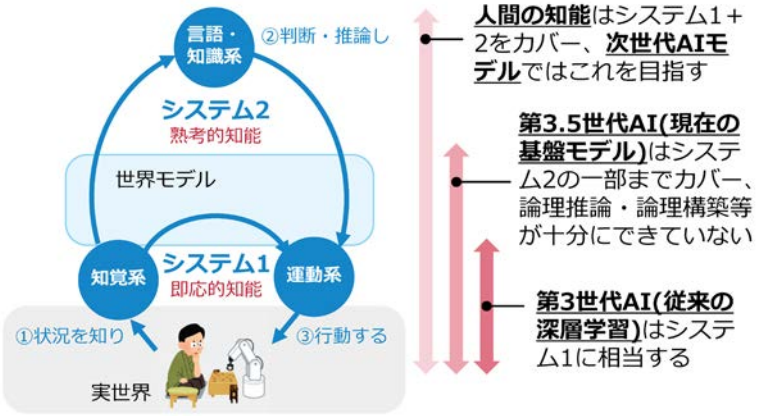


図 B–1–2　二重過程理論と AI モデル

国内でも松尾豊が「知能の2階建てアーキテクチャ」というモデルを発表している[45)]。システム1に相当する1階部分を動物OS、システム2に相当する2階部分を言語アプリと呼び、その二つをつなぐところに世界モデルを位置付けた。国内の人工知能学会全国大会において、世界モデルやシステム1＋2融合に関する研究は特に活発な発表・議論が行われているトピックになっている。

（4）脳情報処理の理解の進展

計算論的神経科学（Computational Neuroscience）は、脳を情報処理システムとして捉えて脳の機能を調べる研究分野である[11]。最近10～20年の間に、脳の機能・活動を知るための計測・理解技術が大きく発展した。特に、神経活動に伴う血管中の血液の流れや酸素代謝の変化を計測・可視化するfMRI（Functional Magnetic Resonance Imaging：機能的磁気共鳴画像法）が、人間を対象に非侵襲で脳の活動を調べることができる計測法として発展した。fMRIによって、人間の行動（心の状態を含む）と脳の活動の同時計測が可能になり、どのような行動や心の状態のときに、脳のどの部分が深く関わっているのか（脳機能マッピング）が調べられるようになった。さらに、マッピングだけでなく、脳の情報処理のモデル化や詳細な比較分析・関係分析などを可能にする手法（ブレインデコーディング、モデルベース解析、Voxel Based Morphometry、拡張テンソル画像、安静時fMRIなど）の発展によって、脳の情報処理についての理解が進展した。中でもブレインデコーディングは、fMRIなどによって計測された人間の脳の活動データを、機械学習の手法を用いて解析することで、人間の心の状態を解読しようとする技術である。脳に想起されたものを、1,000を超える数のカテゴリーと対応付けたり、対象物（名詞）やその動作（動詞）だけでなく、それらの印象（形容詞）をデコードしたりすることも可能になりつつある[46), 47)]。

脳情報処理への関心が触媒となっている機械学習の手法は大変多い。深層学習はもとより、その源流であるニューラルネットワーク、誤差逆伝搬法、自己組織化マップ、表現学習、独立成分解析、強化学習、情報幾何などがある。例えば、強化学習は、ドーパミン神経細胞の報酬予測誤差仮説によって、AI研究における強化学習と脳の強化学習とが強く結び付いている[48), 49)]。AI研究における強化学習は、学習主体が、ある状態で、ある行動をしたとき、その結果に応じた報酬が得られるタイプの問題を扱い、より多くの報酬が得られるように行動を決定する意思決定方策を、行動と報酬の受け取りを重ねながら学習していく機械学習アルゴリズムである。一方、中脳にあるドーパミン神経細胞は、報酬予測誤差（実際に得た報酬量と予測された報酬量との誤差）に基づいてドーパミンを放出し、これが大脳基底核に運ばれることで、脳における強化学習の学

11　この分野の詳しい動向は報告書[7)]の2.1.7にまとめた。

習信号として働くということが分かってきた。また、脳における学習・意思決定のプロセスにはモデルフリー型とモデルベース型があり、モデルフリー型では、刺激と反応の関係性を報酬の程度・確率に直結した形で学習し、モデルベース型では、刺激や反応の間の関係性を状態遷移などの内部モデルとして学習する。モデルフリー型は上述の大脳基底核、モデルベース型は大脳新皮質、特に前頭前野が重要な役割を果たしていると見られている。

2023年初旬の大規模言語モデルなどの生成AIの発展を受け、世界的に著名なAI研究者と神経科学者のグループが、次世代のAI研究の方向性として「NeuroAI」を示した論文を発行した。「NeuroAI」とは「神経科学とAIの交点に位置する新興分野」であるとし、昨今のAIはチューリングテストをパスしかけている（つまり人間と見まがう対話が可能になりつつある）ものの、ロボティクスにおける感覚運動能力は単純な動物にも劣ることを指摘し、動物のように実世界で振る舞えることを試す「Embodied Turing Test（身体化されたチューリングテスト）」をNeuroAIのグランドチャレンジとして提唱している[50)]。

（5）予測符号化理論に基づくAIモデル設計への取り組み

予測符号化理論は、発達科学や認知発達ロボティクス[12]の分野で研究開発が進んでいるもので、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・推論などの認知機能を発達させていく過程をモデル化しようというものである。予測誤差最小化原理（自由エネルギー原理）によって、さまざまな認知発達を統一的に説明することが試みられている。すなわち、現時刻・空間の信号から、将来や未知空間の信号を予測できるように、その対応関係（内部モデル）を学習するメカニズムであり、身体や環境からの感覚信号と、脳が内部モデルをもとにトップダウンに予測する感覚信号との誤差を最小化するように、内部モデルを更新したり、環境に働きかけるような運動を実行したりする（2.1.6の図2–10）。

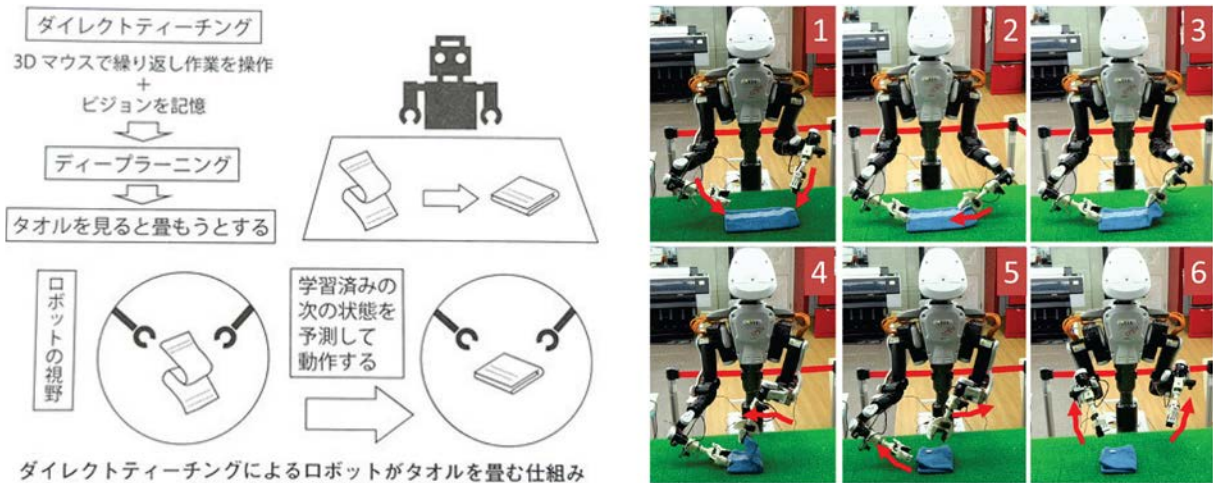


図 B–1–3　深層予測学習によるロボット動作生成[53)]

人工知能学会誌38巻6号（2023年11月）で「自由エネルギー原理とAI」（論文10編）[51)]、日本ロボット学会誌40巻9号（2022年11月）で「予測に基づくロボットの動作学習」（論文8編）[52)] が特集されるなど、AI・ロボット分野で注目が高まっている。予測誤差最小化原理（自由エネルギー原理）を応用したスマート

12　この分野の詳しい動向は報告書[7)] の2.1.8にまとめた。

ロボットやシミュレーターなども開発されている。尾形哲也らが開発したタオル畳みロボット（図B–1–3）では、人間によるタオル畳み操作を手本として、それを模倣する動作を生成しつつ、動作を実行した結果の事前の予測と、実際に実行した結果をカメラで観測して比較し、その差異（予測誤差）から学習することが行われている。予測に使われるモデルは、模倣と予測誤差からの学習によって構築・修正される[53]。また、長井志江らが開発した自閉スペクトラム症（ASD）知覚体験シミュレーター[54]は、予測符号化理論に基づいてASD者の知覚体験を疑似的に再現してみせたことで、ASDの理解・支援に結び付けた。

このようなシステムへの実装が行われているものの、知能や認知発達過程のモデル化という目標に対してはまだ限定的で道半ばである。現在の基盤モデルが抱える問題点の克服にどうつなげられるかも課題である。なお、言語の獲得・創発という面で、谷口忠大が予測符号化理論をベースとしつつ、二重過程理論に重ね合わせたモデリングを試みている[18]のが興味深い。

（6）人間や環境との関係性に基づくAIモデル設計への取り組み

AI単体でより高度な処理機能を持たせる方向性の他に、AIと人間や環境との関係性という面でより高度な処理機能を実現しようという考え方がある。二つの研究事例を取り上げる。

一つは、対話システムの観点から、コモングラウンドが必要だという指摘である。西田豊明は、コモングラウンドを、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のことと定義した[55]。ChatGPTなどの生成AIは、ユーザーからの自然言語入力（プロンプト）に応答を返すが、それは確率モデルに基づく予測であって、生成AIとユーザーとの間にコモングラウンドは形成されていない。コモングラウンドの役割・必要性は認識されつつあるが、どのようにしてこれを形成するかは今後の課題である。

もう一つは、環境の側にも知的な機能を持たせようという、アフォーダンス的なアプローチである。その一例として、三宅陽一郎が提案したMCS-AI動的連携モデル[56]が挙げられる。このモデルでは、ゲーム空間において、メタAI（M）、キャラクターAI（C）、スパーシャルAI（S）[13]という3種類のAIが動的に連携して動作する。メタAIはゲーム全体を監視し、キャラクターAIはノンプレーヤーキャラクターの動作を状況認識に基づいて決定し、スパーシャルAIは空間を認識する。実際にゲーム開発に適用されており、2020年度人工知能学会論文賞も受賞した。

（7）基盤モデルを用いたロボット制御への取り組み

現在の基盤モデルの問題点として、実世界操作が苦手であることが指摘されており、その通りである一方で、ロボットの動作制御・動作生成のために基盤モデルやトランスフォーマー型モデルを用いる試みが進められている。その先進的な研究開発事例として、PaLM-SayCanやRX-1/2/Xが挙げられる。

PaLM-SayCan[57]は、GoogleとEveryday Robotsの共同開発によって2022年8月に発表された。このシステムでは、自然言語による曖昧な要求に対して、その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することが示された。例えば「飲み物をこぼしてしまった。助けてくれる？」という問いかけに対して、スポンジを取ってくるという行動を起こす。PaLM-SayCanを用いたロボットは、自然言語による問いかけを解釈し、事前に定義されたスキルセットの中から、その状況で実行可能で、要求に対して有効な行動を選択する。

大規模言語モデル・生成AIは、トランスフォーマーと呼ばれる深層ニューラルネットを用いているが、これに大量の動作データを学習させるロボティクストランスフォーマーの研究開発が急速に立ち上がってきた。そのきっかけとなったのはPaLM-SayCanと同じくGoogleとEveryday Robotsから、2022年12月に発表されたRT-1（Robotics Transformer 1）[58]である。RT-1では、Everyday Robotsのロボット実機13台で

13　論文[56]中ではナビゲーションAIと名付けられているが、その後、スパーシャルAIと名称が変更された。

17カ月にわたって、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して、大規模学習が実施された。その結果、700種類のタスクで97%の成功率を達成した。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が高められた先進事例である。

さらにRT-1を発展させたRT-2が、2023年7月にGoogle DeepMindから発表された[59]。 RT-2では、RT-1の大規模学習済みトランスフォーマーモデルをベースに、さらにウェブ上のテキストと画像でトレーニングして構築されたVLA（Vision-Language-Action）モデルを用いる。 RT-1を含めて従来は、事前学習済みの物体は、操作の対象物として認識できるが、未学習の物体は操作対象にできない。しかし、ウェブから獲得した知識によって、その物体に関する知識を得ることで、必要な動作を生成することが可能になる。

さらに、2023年10月には、RT-1、RT-2を発展させたRT-Xの研究開発を推進するOpen X-Embodimentが発表された[60]。 Google DeepMindを中心とした世界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である。22種類のロボットハードウェアから集めた実機での動作データで、527種類のスキル、100万種類のエピソードを含んでいるという。 RT-1モデルをベースとしたRT-1-Xと、RT-2モデルをベースとしたRT-2-Xが開発されている。

このような先進的な研究開発事例があるものの、ロボットによる実世界操作は、ロボットの個体差や実世界状況の多様性が大きいことに対して、現在の基盤モデルの枠組みから、どのような拡張・発展が必要になるのかは今後の課題であろう。

B.1.2　AIリスクへの対処のための技術開発

（1）AI応用システム開発における品質管理

近年、機械学習技術を用いたAI応用システムが開発されるようになったが、従来のプログラミングによる開発は演繹的な作り方であるのに対して、機械学習を用いた開発は帰納的な作り方になる（**コラム2**参照）。すなわち、開発パラダイムが異なるので、従来の開発方法論が通用せず、AIシステム開発では安全性・信頼性の確保のために新しい方法論・技術群が必要になった。そこで、2018年頃から新たに機械学習工学[24]（AIソフトウェア工学[25]）と呼ばれる研究分野が立ち上がり、活発な研究開発が行われている[14]。具体的には、機械学習の品質管理・テスト手法、公平性評価法、プライバシー保護の技術、説明可能AI、脆弱性対策などの技術開発が行われている。

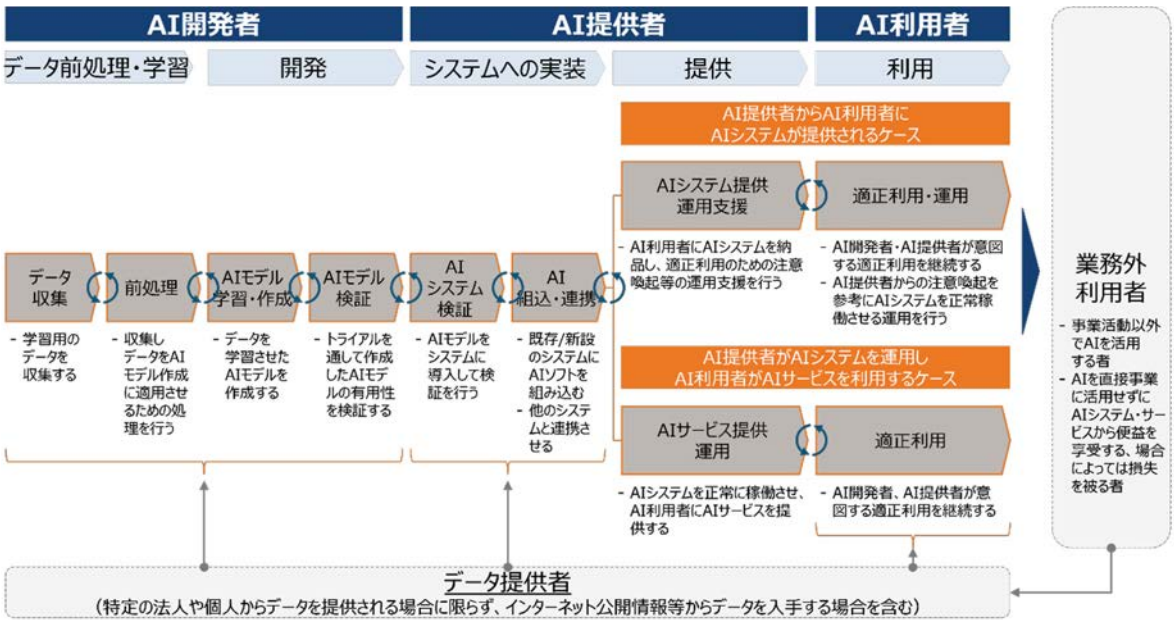
また、機械学習応用システムの開発ガイドラインも整備され、国内では特に、QA4AIコンソーシアムの「AIプロダクト品質保証ガイドライン（QA4AIガイドライン）」[15]（2019年5月初版、その後毎年改版、2024年1月最新版）と産業技術総合研究所の「機械学習品質マネジメントガイドライン（AIQMガイドライン）」[16]（2020年6月第1版、2021年7月第2版、2022年8月第3版、2023年12月第4版）が、詳細で実践的なものとしてよく知られている。QA4AIガイドラインは、五つの軸でチェックリストがまとめられており、また、自動運転、生成系システム、産業用システム他、ドメイン別のガイドラインも作成されている。AIQMガイドラインは、機械学習の品質や品質管理の考え方から整理されている。

また、2024年1月に総務省・経済産業省から「AI事業者ガイドライン案」が公表された[17]。日本政府からはこれまでに、「人間中心のAI社会原則」（内閣府、2019年）に基づき、「国際的な議論のためのAI開発ガイドライン案」（総務省、2017年）、「AI利活用ガイドライン～AI利活用のためのプラクティカルリファレンス～」（総務省、2019年）、「AI原則実践のためのガバナンス・ガイドライン」（経済産業省、2022年）が出されて

14　この分野の詳しい動向は報告書[7]の2.1.4にまとめた。
15　https://www.qa4ai.jp/
16　https://www.digiarc.aist.go.jp/publication/aiqm/
17　https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240119_report.html

いたが、それらを統合し、生成AIをはじめとする最新動向を考慮して見直したものが「AI事業者ガイドライン案」である。図B–1–4は同ガイドラインから引用した図であるが、AIライフサイクルを踏まえ、AI開発者、AI提供者、AI利用者が留意すべき点がまとめられている。

以上は、2.1.6（2）や図2–11の分類において、主にケース2、すなわち、生成AIのアプリケーションを開発するケースに該当する。一方、ChatGPTやGitHub Copilotの登場以降、ソフトウェア開発の現場では、ケース2だけでなく、ケース3のような形での生成AI活用が急速に進んだ。つまり、ソフトウェア開発プロセスにおける各ステップ（要件明確化、アーキテクチャー提案、ソースコード生成、テストプログラム生成、デバッグなど）で、生成AIがいわば助手のように活用されるようになってきた[18]。このようなケース3における品質管理も、新たな課題になる。



図B–1–4　AIライフサイクルと事業活動を担う主体

（「AI事業者ガイドライン案」から引用）

（2）生成AIの出力品質の確保

チャットボット関連の事件として、2016年3月23日に、MicrosoftのTayがTwitter上で公開されたが、悪意を持ったユーザーとの対話によって、ごく短時間で差別主義的思想に染まってしまい、開始後16時間で強制終了となったことがよく知られている。このような問題を回避し、対話型生成AIがより適切な応答を返せるようにするため、複数通りの対策が取られている。

まず、学習データの選別が行われる。不適切な内容のデータは取り除き、学習に使わないことでリスクを低減するものである。

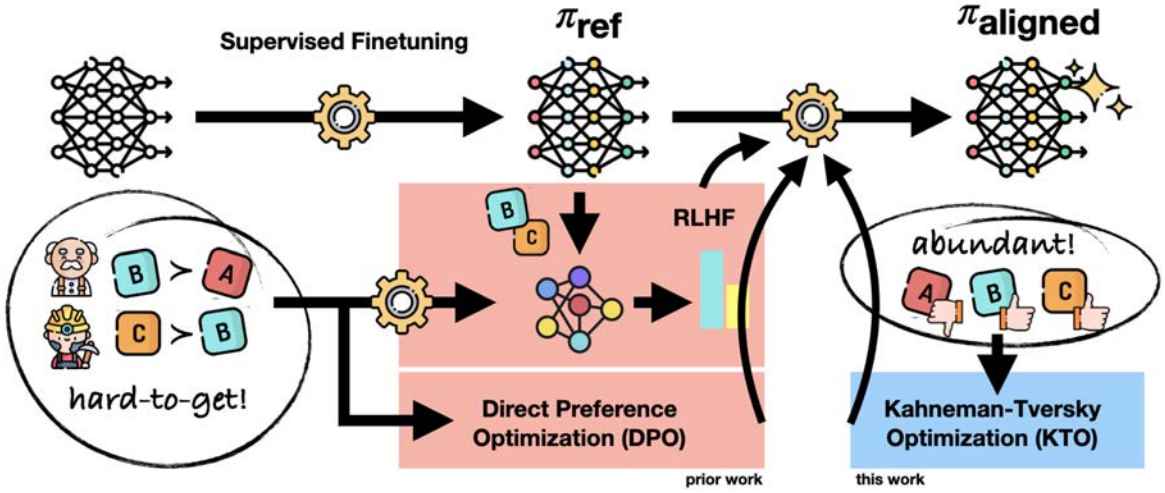
また、最終出力に対して、フィルタリングをかけるContent Moderationも行われている。不適切な表現（性

18　このような事例は、例えば以下のようなイベントなどで紹介されている。
「2023年度JEITA ソフトウェアエンジニアリング技術ワークショップ ～生成AIによってソフトウェアエンジニアリングはどう変貌するか～」（2024年2月9日開催）https://home.jeita.or.jp/system/seminar/240110.html
「GLOCOM六本木会議オンライン#73 ChatGPTなどの生成AIを用いたソフトウェア開発の最前線」（2024年2月5日）https://roppongi-kaigi.org/event/2852/

的/暴力/ヘイトなど）を含む応答を検出してブロックするというものである。

さらに、適切な応答を返せるようにモデルをファインチューニングする技術（インストラクションチューニング）が開発された。特に有名なのはChatGPTに導入されたRLHF（Reinforcement Learning from Human Feedback）[14]、すなわち、人間のフィードバックを用いた強化学習によって基盤モデルをチューニングするというものである。 RLHFでは、まず、プロンプトと望ましい応答のペアを追加学習する。次に、応答文のスコア付けモデル（真実性、無害性、有益性に基づいてスコア付け）を人間による比較評価結果（選好データセット）から学習する。最後に、学習したスコア付けモデルを用いて、大量のプロンプトで強化学習を行い、基盤モデルをチューニングする、ということが行われる。

インストラクションチューニングの手法としては、RLHF以外にも、強化学習を用いず、選好データセットをもとにモデルを直接チューニングするDPO（Direct Preference Optimization）[93] や、選好データセットではなく、プロンプトに対する応答に「良い／悪い」の評価ラベルを付けたものをもとにモデルをチューニングするKTO（Kahneman-Taversky Optimization）[94] などが考案されている（図B–1–5）。



図B–1–5　RLHFとDPOとKTOの違い[94]

（3）フェイク対策

生成AI技術の高精度化に伴い、フェイク問題が大きな懸念事項になっている。ここでは、その対策の動向について述べる。

まず、ソーシャルメディア上での情報伝播の傾向や、そこで起きている炎上、フェイクニュース、エコーチェンバー、二極化などの現象を把握・分析することが、フェイク対策のための基礎的研究となる。フェイクニュースへの対抗としては、発信された情報が客観的事実に基づくものなのかを調査し、その情報の正確さを評価・公表するファクトチェック[61] という取り組みが立ち上がっている。ファクトチェックを行う団体として比較的早期に立ち上がった米国のSnopes（1994年～）やPolitiFact（2007年～）がよく知られている。この動きは世界的に広がっており、日本では2017年にFactCheck Initiative Japan（FIJ）、2022年にJapan Fact-check Center（JFC）が発足した。2015年には国際的に認証されたファクトチェック団体から成るInternational Fact-Checking Network（IFCN）も発足した。

ただし、大量に発信される情報を迅速にチェックするには人手では限界がある。そこで、コンピューター処理によってフェイクニュースの検出を支援する試みが進められている。フェイクニュース検出は、次のような四つの面から試みられている[62]。一つ目は知識ベース検出方式で、従来の人手によるファクトチェックを強化するように、クラウドソーシング的な仕組みを使って専門家集団に検証してもらったり、あらかじめ蓄積された

知識ベースと自動照合したりする取り組みがある。二つ目はスタイルベース検出方式で、誤解を生みやすい見出し表現や欺くことを意図したような言葉使いなどに着目する。三つ目は伝播ベース検出方式で、情報拡散のパターン（情報伝播のグラフ構造やスピードなど）に着目する。例えば、フェイクニュースは通常ニュースよりも速く遠くまで伝わる傾向があることが知られている。四つ目は情報源ベース検出方式で、ニュースの出典・情報源の信頼性やその拡散者の関係などから判断する。

社会環境・文脈などによって真偽の捉え方が変わるし、科学的発見によって真理の理解が変わることもあり、真偽が定められない言説も多いため、最終的には人による判断が不可欠だが、上に示したような技術は怪しいニュース・情報を迅速に絞り込むのに有効である。また、1件のニュース単独で真偽判定するよりも、複数の情報の間の関係比較や整合性判断、および、複数の視点からの複合的なチェックを行う方がより確かな判断が可能になる[23)]。

さらに、フェイク動画・フェイク画像・フェイク音声などの判定については、オリジナルの動画・画像・音声から改ざんされていないか、当事者が実際に発話や行動をしていない虚偽の動画・画像・音声ではないか、といったことを動画・画像・音声の特徴分析、ニューラルネットワーク・機械学習を用いて判定したり、改ざんの箇所や方法を特定したりといった技術が開発されている[63), 64), 65)]。例えば、不自然なまばたきの仕方、不自然な頭部の動きや目の色、映像から読み取れる人の脈拍数、映像のピクセル強度のわずかな変化、照明や影などの物理的特性の不自然さ、日付・時刻・場所と天気の整合などが手掛かりになる。フェイクの検出技術とフェイクの作成法はいたちごっこだとも言えるが、人の目・耳では見分けが付かないレベルのフェイク動画・画像・音声が作れてしまう事態において、コンピューターによる分析は不可欠である。さらに、動画・画像・音声の内容解析とは別に、電子透かしやブロックチェーンを使って出所や経路を追跡することで、改ざんが入り込むことを防ぐという方法もある。

フェイクは新種のサイバー攻撃として用いられ、社会混乱も引き起こすことから、安全保障上も対策が求められる。これに対して、米国は早い時期から研究投資を行っており、特に米国国防高等研究計画局（DARPA）は、画像・動画の改ざん検知を行うMedia Forensics[19]（MediFor、2015年開始）プロジェクトや、その後継で、画像・動画の意味的不整合やフェイクの検知を行うSemantic Forensics[20]（SemaFor、2021年開始）プロジェクトを推進している 。日本では、2020年度にJST CREST「信頼されるAIシステム」に採択された「インフォデミックを克服するソーシャル情報基盤技術」（CREST FakeMedia）[21]が、フェイク問題に本格的に取り組むプロジェクトとして注目される。この活動を推進する拠点として、国立情報学研究所にシンセティックメディア国際研究センター（SynMedia Center）[22]が設立された。同センターからは、フェイク顔映像を自動判定するプログラムSYNTHETIQ VISIONがリリースされ、ライセンス事業化も行われている。

19 https://www.darpa.mil/program/media-forensics
20 https://www.darpa.mil/program/semantic-forensics
21 https://research.nii.ac.jp/~iechizen/crest/index.html
22 https://research.nii.ac.jp/~iechizen/synmediacenter/index.html

B.2　AIガバナンスに資する研究体制に関する国内外動向

本節では、研究開発課題②「AIリスクへの対処」と研究開発課題④「人・AI共生社会の在り方」に関連の深いトピックとして、過去1、2年のAIガバナンスの動向と、それを支える研究・実践を行う国内外の機関・拠点の現状を概観し、最後に日本における課題と検討すべき方向性を述べる。

B.2.1　2024年初頭における国際的なAIガバナンスの動向

AIガバナンスとは、AIがもたらすさまざまな倫理的・法的・社会的課題（ELSI）に対処しつつもAI技術によるインパクトを最大化するための、技術・組織・社会システムの設計と運用を指す[23]。とりわけ政策的なルール形成に関わるAIガバナンスは、2010年代半以降、原則レベルの議論が進展し、2019年に採択されたOECDにおける「AI原則」を経て、ガイドラインや法規制案といった実践レベルの議論へと進展してきた[24]。そのような中、2022年末に登場したChatGPTは社会的に大きなインパクトを与え、生成AI・基盤モデルに対する規制の必要性への関心が急速に高まり、世界各国・各国際機関における主要アジェンダとなった（**表B–2–1**）。今後も、技術の進展による新たな論点を加えながら、AIガバナンスの議論はますます加熱していくと考えられる。

表B–2–1　2024年初頭における国際的なAIガバナンスの主要動向

EU	2023年6月、欧州議会は欧州AI法案の修正案を採択。2023年12月、政治的合意。
米国	2023年7月、大手AI企業7社と自主コミットメントに合意。 2023年10月、AI安全に係る大統領令を発出。11月、AI Safety Instituteの設立発表。
英国	2023年11月、AI安全サミット開催、ブレッチリー宣言発表。AI Safety Instituteの設立発表。
日本	2023年12月、新AI事業者ガイドライン案発表。2024年2月、AI Safety Instituteの設立準備中。
中国	2023年8月、生成AIに関する規制法施行。 2023年10月、グローバルAIガバナンスイニシアチブ発表。
G7	2023年5月、「広島AIプロセス」の開始を宣言、2023年10月、高度なAIシステムを開発する組織向けの国際指針、国際行動規範などを発表。
国連	2023年10月、AIに関するハイレベル諮問機関設立。2023年12月、中間報告書を公表。
欧州評議会	AI枠組み条約の交渉が進展中。2024年春妥結を目指す。

B.2.2　AIガバナンスにつながる研究・実践を行う国外の機関・拠点

国際的なルール形成の担い手は各国の政策立案者だけではない。そこには、人文・社会科学を含む多くの分野の研究者からの知見が提供され、議論形成に不可欠なものとして活用されている。さらに、AIの影響を被る市民社会からの声を集め、届ける役割をアカデミア研究者が担うことも多い。

AIに関するルール形成は、先述のように、原則フェーズの議論から実践フェーズの議論へと進展してきた

23　参考：経済産業省「我が国のAIガバナンスの在り方 ver. 1.1」ではAIガバナンスの定義として「AIの利活用によって生じるリスクをステークホルダーにとって受容可能な水準で管理しつつ、そこからもたらされる正のインパクトを最大化することを目的とする、ステークホルダーによる技術的、組織的、及び社会的システムの設計及び運用」を採用している。
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/2021070901_report.html

24　関連動向は報告書[7)]の2.1.9を参照。

面はある一方、急速な技術進歩を背景に原則を問い直す必要も生じる[25]。そのため、今般のAIガバナンスにおいては、具体的な標準化・評価指標などにつながる具体的な研究・実践やAIのリスクを低減・制御する技術的研究（**第3章**の研究開発課題②「AIリスクへの対処」）のみならず、AIの社会への長期的影響や、AIと人間・社会のあるべき関係を問い直す射程の長い研究（研究開発課題④「人・AI共生社会の在り方」）の知見も求められている。

欧米には、こうしたAIガバナンスに不可欠な学術知を生み出す研究や実践、提言活動を行う拠点や組織が存在する（**表B–2–2**に例を示す）。こうした拠点・組織は、AIの社会への長期的影響、AIと人間・社会のあるべき関係についての研究や、政策提言活動などを通じて、現下の国際・国内AIガバナンスへの重要な論点を提供している。

表B–2–2　国際的なAIガバナンスを支える海外の研究拠点・組織の例

大学・公的研究機関における拠点の例	米 ニューヨーク大学 AI Now Institute	2017年設立のAI倫理系の研究センター。巨大IT企業への権力集中に対する問題意識などからの多くの提言（ポリシーブリーフなど）を出している。
	米 スタンフォード大学 Institute for Human-Centered AI（HAI）	米国アカデミアのAI研究の一大拠点であるスタンフォード大学にて、2019年始動。最先端のAI研究とともに、AI倫理教育などにも力を入れる。アカデミア向けAI研究資源整備に関する提言など、米国のAI政策に影響力を持つ。
	米 ジョージタウン大学 Center for Security and Emerging Technology	中国のAI開発の動向などについて精力的な情報発信を行う他、米国のAI政策に対する政策提言も行う。
	米 ミシガン大学 公共政策大学院	同大学院のScience, Technology, and Public Policy Programのメンバーが、2022年に大規模言語モデル（LLM）の影響に関する分析レポートを公開。
	英 アラン・チューリング研究所	2015年設立のAIとデータサイエンスの国立研究所。「データ倫理グループ」などに倫理学者・哲学者が在籍し、倫理面の研究や政策提言を積極的に行う（**第4章**の**コラム5**参照）。
	英 オックスフォード大学 The Future of Humanity Institute	2005年設立の学際的な研究センター。民主主義や社会的公正、ジェンダーなどの人文・社会科学系のアプローチも含む形で、拠点の研究プログラムが再編成された。
	英 ケンブリッジ大学 Centre for the Study of Existential Risk	Jaan Tallinn氏、Hew Price氏らにより2012年設立。Existential Risk（人類の存亡リスク）研究の草分け。バイオ、地球環境、極端テクノロジー、AIなどのリスクの認識・分析、共有・合意、問題解決のコミュニティー形成に取り組む。
	蘭 アムステルダム大学 RPA Human(e) AI	2019年開設。人間中心のAI活用を促進するための研究を行う。学部2、3年生向けに、Human(e) AI Bachelor courseを提供。
	独 ミュンヘン工科大学 Institute for Ethics in Artificial Intelligence（IEAI）	Science, Technology and Society（STS）の世界的拠点であるミュンヘン工科大学にて、AI倫理に関する学際研究を行う組織として2019年設立。

25　江間有沙氏はこの状況を「原則も実践も」と表現している。参考：「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」における江間有沙氏資料「AIガバナンスのPrinciples and Practices」
https://www.soumu.go.jp/main_content/000913994.pdf

	蘭 エラスムス・ロッテルダム大学 Erasmus Centre for Data Analytics（ECDA）	2018年開設。同大学が設定するSocietal Impact of AI（AiPact）イニシアチブを中心的に推進。
民間・非営利組織・学会・コンソーシアム（ごく一例）	米 Future of Life Institute	2014年3月設立。2017年1月「アシロマAI原則」を始め、AIガバナンスに影響力を持つ非営利組織の一つ。2023年3月、GPT-4を超える大規模AIシステムの開発モラトリアムに関する公開書簡を公開。研究助成事業も手掛ける。
	米 Partnership on AI	2016年9月にGoogle、Meta、Amazonなどが参画して、マルチセクターで責任あるAIを推進することを目的に設立された非営利組織。
	米 Center for AI Safety	2022年設立の非営利組織。2023年5月にはAIによる絶滅リスクに関する声明を出し、トップAI研究者らが署名。
	英 Ada Lovelace Institute	Nuffield財団が2018年に設立した非営利研究所。データやAIの公正な活用などを目指して活動。
	英 Centre for the Governance of AI	オックスフォード大学を母体とし2021年に非営利組織として独立。AIガバナンスに関する研究・提言活動を行う。
	米 The AI Alliance	2023年12月にIBMとMetaが発起人となり、50を超える設立メンバーとともに、オープンで安全な、責任あるAIの推進に向けた国際コミュニティーとして発足。日本からは東京大学、慶應義塾大学、Sakana AI、SB Intuitions、ソニーグループ、フェンリルが設立メンバーとして参加。

B.2.3　2023年の特筆すべき政策的取り組み

AIガバナンスへの関心の高まりを受け、2023年に入ってから各国は研究体制のさらなる強化に取り組んでいる。例えば2023年11月に英国で開催されたAI安全サミットでは、英国と米国はともにAI安全研究所（AI Safety Institute）の設立を発表した。

- 英国 AI Safety Institute：11月2日に英国政府が発足を発表。（1）先端AIシステムの評価の開発と実施、（2）基盤AIの安全研究の推進、（3）情報交換の推進（政府当局、海外機関、企業などを含む）、を三つの中核機能とする。
- 米国 AI Safety Institute（US AISI）：商務省・NIST（国立標準技術研究所）内に設置し、NISTが開発したAIリスクマネジメントフレームワーク（AI RMF）を実装すべく、ガイドライン、ツール、ベンチマーク、ベストプラクティスなどの作成、レッドチームの評価、コンテンツ認証、透かし、有害アルゴリズムの特定などに係る技術ガイダンスの開発を行うとする。2024年2月、NISTは200超の企業・大学・NPOなどからなる「AI Safety Instituteコンソーシアム」の始動を発表した。

その他、英米は相次いで新しいプログラムを始動している。

- 英国 Responsible AI UK（RAI UK）：2023年6月始動。5年間で3100万ポンド（～56億円）をUKRIがファンド。サザンプトン大学が中核となり、産業界も含むAIの国際的なエコシステム形成、研究開発プログラムの実施、スキル形成、パブリック&政策エンゲージメントを掲げる。
- 英国 Bridging Responsible AI Divides（BRAID）：2023年6月始動。芸術・人文科学研究会議（AHRC）にて、3年間、850万ポンド（～15億円）をファンド。エジンバラ大学を中心にAda Lovelace 研究所およびBBCの協力のもと実施。責任あるAIに向けてSTEM領域と芸術・人文学の接続を掲げている。

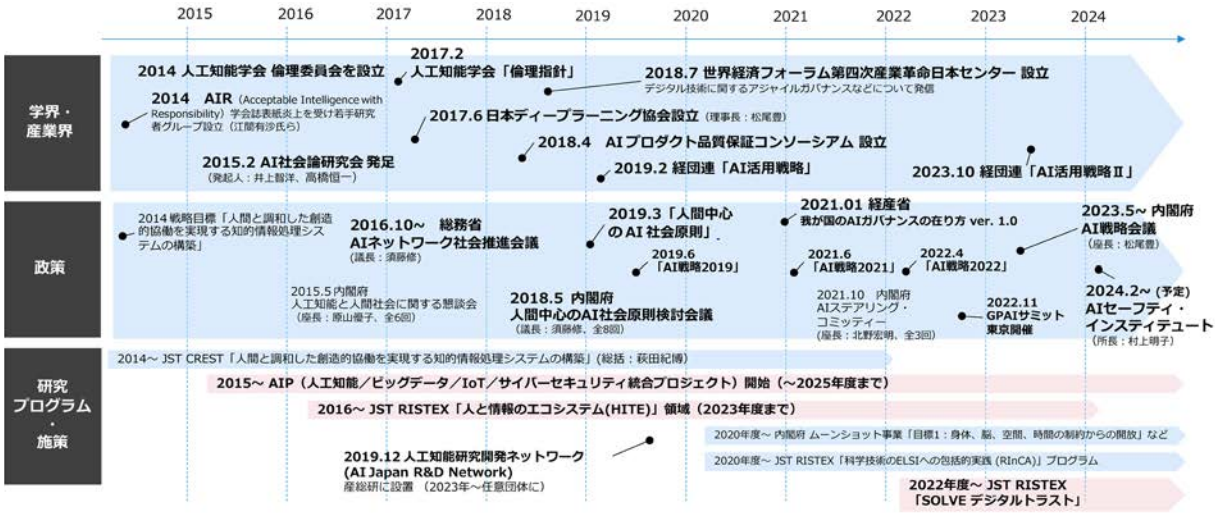
・米国TRAILS（Institute for Trustworthy AI in Law & Society）：全米に設置された20数拠点から成る国家AI研究所の一つとして、2023年5月にメリーランド大学を中心に設立。 NSFより5年間、2000万ドルのファンドを受ける（第4章のコラム5も参照）。

B.2.4　日本のこれまでの動き

日本でも2010年代中頃より、政策・行政、産業界、学界がさまざまな取り組みを行ってきた（図B–2–1）。2016年にG7伊勢志摩サミット議長国として提案したAI研究開発のガイドライン案は、AIガバナンスの国際議論の先駆けの一つであった。2017年には人工知能学会の倫理指針、総務省の「国際的な議論のためのAI開発ガイドライン案」が公表された。その後、2019年3月に内閣府から「人間中心のAI社会原則」が公表され、それらを踏まえ、同年5月にはOECD AI原則が採択されている。 AI分野のルール形成への取り組みは、国際標準化活動SC42やGPAIが中心的な活動母体となり、研究コミュニティーや政策立案者との連携を含む体制・枠組みもある程度機能している。国際的なポジションとしても、GPAIやSC42などにおいて日本は重要な役割を担い、善戦していると言える。

こうした蓄積に加え、（1）AIにとどまらない新興技術のELSIや科学技術ガバナンスに取り組む大学拠点（「ELSIセンター」など）の設立、（2）AIガバナンスに取り組む業界団体の登場などにより、2024年初頭現在、国内のAIガバナンス/AI ELSIに関わる多くの拠点や取り組みが存在している（表B–2–3）。

また、日本においても英米と同様に「AIセーフティ・インスティテュート（Japan AI Safety Institute：AISI）」を設立し、国内のAI研究拠点と連携しながら、AI安全性に関する評価や技術開発を行う予定である[26]。



図B–2–1　　日本における、AIガバナンスに向けた産官学のさまざまな取り組み

26　AIセーフティ・インスティテュート（AISI）https://aisi.go.jp/

表B–2–3　AIガバナンスに関連するテーマに取り組む国内の研究拠点・組織の例

種別	拠点・組織名	活動の例
大学・公的研究機関における拠点の例	理化学研究所AIPセンター「社会における人工知能研究グループ」	人工知能の進展が人間社会に及ぼす影響の分析と対策の研究など。
	産業技術総合研究所 情報・人間工学領域	AI国際標準に係る中核機関としての取り組み（ISO/IEC JTC 1/SC 42対応、NISTとの連携など）、機械学習品質マネジメントガイドラインの発行など。
	東京大学未来ビジョン研究センター	AIやデータガバナンスに関する研究活動と、それに基づく各種の政策提言（例：2023年3月「AIガバナンス協調への道筋：G7サミットに向けた政策提言」）。
	東京大学 Beyond AI 研究推進機構 B'AI Global Forum	東京大学Beyond AI研究推進機構のプロジェクトとして、2020年9月設立。 AI時代のジェンダー平等社会、マイノリティーの権利保障のための規範・倫理・実践研究。
	大阪大学ELSIセンター	生成AIのELSIの論点をまとめたELSI noteを公開。
	中央大学ELSIセンター	2021年設立。 AI・セキュリティーに特化した研究活動を柱に、須藤修センター長を中心に国際活動にも注力。
	立教大学 人工知能科学研究科	2020年開設。社会科学×AIやAI ELSIをカリキュラムに組み込んでいる。
	北海道大学人間知×脳×AI研究教育センター（CHAIN）	AIと哲学・倫理学にまたがる研究活動の他、2023年度にはリカレント教育として「AI倫理コース」開講。
	慶應義塾大学グローバルリサーチインスティテュート（KGRI）	サイバー空間にまつわる社会システムや法政策の研究拠点であるサイバーフィジカル・サステナビリティ・センターなどを擁する。
民間・非営利組織・学会・コンソーシアム	AIガバナンス協会	2023年10月 Robust Intelligence社共同創業者の大柴行人氏が発起人となり設立。 NECやNTTデータなど23社が参加。
	一般社団法人 日本ディープラーニング協会	ディープラーニング技術を日本の産業競争力につなげることを目的として2017年に設立。「AIガバナンスとその評価」研究会では、AIの産業構造を踏まえたAIガバナンス・エコシステムに関する報告書を発行。
	AIプロダクト品質保証（QA4AI）コンソーシアム	産学のメンバーにより2018年設立。 AI製品の品質保証のためのガイドラインを発行。
	一般社団法人 京都哲学研究所	2023年7月、NTT株式会社澤田純会長と京都大学出口康夫教授を共同代表理事として設立。価値多層社会を掲げ哲学を軸とした研究活動や、2025年の「京都会議」開催を予定。
	国際会議 Japan AI Alignment Conference 2023	2023年1月にConjecture社とアラヤ社が共催したAIアライメントのワークショップ。国内外から約60人の研究者が参加。（「AIアライメントとは」については**付録C**参照）
	オリジネーター・プロファイル（OP）技術研究組合	2022年12月設立。理事長は村井純教授（慶應大学）。インターネット上のニュース記事や広告などの情報コンテンツに発信者情報を紐付ける仕組み構築を目指し、大手新聞社などメディア各社が加入。
	一般社団法人 Generative AI Ja–pan	2024年1月、株式会社ベネッセコーポレーションとウルシステムズ株式会社を共同発起人とし、慶應義塾大学の宮田裕章教授を代表理事として発足。生成AIを主としたAIに関する認定制度の検討、教育活動、AIに関する政策提言などを実施予定。

こうした各拠点・組織とは別に、国の研究プログラムの中でもAIガバナンスにつながる研究の蓄積がなされてきた。中でも、JST社会技術研究開発センター（RISTEX）の「人と情報のエコシステム（HITE）」領域（2016～2023年度）は、「情報技術と人間のなじみがとれている社会を目指すために、情報技術がもたらすメリットと負のリスクを特定し、技術や制度へ反映していく共進化プラットフォームの形成を行う」ことを掲げ、過去7年間に24件の課題を採択し、さまざまな知見・議論の蓄積と、人材育成、コミュニティー形成につなげてきた。研究活動だけでなく、コミュニティーづくりやメディア・コミュニケーション戦略においてもさまざまな試行を行ってきた同領域の活動は、世界的に見てもユニークなものであり、その蓄積を今後のプログラム設計に生かすことが望まれる。

B.2.5　日本の取り組みの課題と、検討すべき方向性

以上見てきたように、AIの分野に関しては、日本は国際社会に先駆けてAIガバナンスの議論を開始し、一定程度先導してきた。また、ここ1、2年の間にも、産学官それぞれから拠点や組織形成の動きがあり、国内外のAIガバナンスへ貢献し始めている[27]。

他方、これまでのCRDSの調査活動から見えてきた課題として、ELSIの同定や社会的インパクトの検討、価値観の言説化といったAIガバナンスの「上流」に位置する活動と、産業界や行政が関与する具体的なルール形成といった「下流」の取り組みが、十分に結び付いていない現状がある[28]。AIガバナンスに資する研究体制に関する課題の明確化を目的に、2023年12月～2024年1月にかけて、AIガバナンスの国際交渉において中心的な役割を果たしてきた4名の専門家にヒアリングを行った。その結果、以下のような課題が同定された。

国際的なルールメイキングへ参加する体制について

- ELSIの方面から国際的なAIガバナンスを支える一部の研究者（以下“コア研究者”と呼称）が存在するが、ごく数名の研究者の俗人的なエフォートに依存している。コア研究者は、一度引き受けると頼られすぎて研究ができなくなる。
- AIガバナンスに関わる専門家同士での情報共有が研究者同士の努力で行われている。コア研究者は多忙すぎて、他の国際会議に出ているコア研究者同士での十分な情報交換ができない。
- これらの活動において、チームとして対応する体制や支援が乏しい。
- 国際交渉に継続的に参画することで信頼される、「顔の見える」国の行政官がごく少数に限定されている。結果として、コア研究者の負担が大きくなる傾向。

マルチセクターでの情報共有の体制

- マルチステークホルダーでの情報集約が、コア研究者による自力の情報収集に依存している。
- 行政の中ではAIに関する会議体が多くあり、省庁間をまたぐ議論も進む。一方、コア研究者にとって行政の横断的な情報共有がなされているのか不透明なケースも多い。各種審議会への専門家招聘もシステマチックには行われていない。
- 海外の一部拠点は市民などマルチステークホルダーからの声を聴ける体制を「強み」として打ち出し、影響力を強めているが、日本ではそうしたことができる主体や場が存在しない。

27　例えば、東京大学未来ビジョン研究センターは、広島AIプロセスや欧州評議会のAI条約交渉に際して、専門家としての参画や提言活動を行っている。ISOなどのAI国際標準については、長らく産業技術総合研究所が中心に対応してきた。

28　このことは、AI分野に限らない、ELSIや新興技術ガバナンスへの日本の取り組み方の課題を別のプロポーザル（CRDS-FY2023-SP-01「科学技術・イノベーションの土壌づくりとしてのELSI/RRI：戦略的な科学技術ガバナンスの実現に向けて」[66)]）にまとめた。

AIガバナンスに資する知識生産

・RISTEX-HITEなどのプログラムにより、AI ELSIに関わる一定の知識生産が行われており、各大学のELSIセンターなどの拠点設立も相次ぐ一方、知見の体系的な蓄積・共有と、継続的な発展性に課題がある。
・AIガバナンスとして必要な、海外ベンチマークや戦略的な政策オプションなどを検討するシンクタンクあるいはその機能が不在。概してこれらは、アカデミックな研究対象にはなりにくい。

人材育成

・過去のプログラムにより、AI ELSIに関わる研究者のコミュニティーが醸成されてきた。人文・社会科学系の各分野にもこの領域で活躍できそうな若手が潜在的に存在するものの、AIガバナンスのコア研究者となる人は少ない。そうした活動は所属機関での評価や、キャリア展望につながりにくい。
・AI技術のことを理解した上でAIガバナンスの担い手となり得る研究者・実務者が足りない。

こうした課題を解決するためには、**第4章**で述べた研究エコシステムの要素のうち、

f. 研究エコシステムのハブ機能を担う組織の設置
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. 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
h. 研究エコシステムを支える人材の確保・育成

が重要となる。AIガバナンスの観点からこれらの実現方策を考えると、以下のような考え方や仕組みがポイントと考えられる。

（1）**AIガバナンスに資する研究・実践のハブとなる機関の設置：**fに示した通り、**表B–2–3**に挙げた多くの拠点・組織の知見を集約・蓄積し、利用可能にするための実践を担うハブとなる拠点機能が求められる。①各研究機関・大学などが連携したバーチャル拠点の設置、②特定の研究機関・大学などへの実拠点の設置、の2パターンが考えられるが、いずれの場合もホワイトペーパーやポリシーブリーフの作成、ステークホルダー間調整などの実務を担う事務局機能が必要となる。また、これから新設されるAIセーフティ・インスティテュートなどとの連携や、AI戦略会議などの重要な会議体や各施策を執行する行政などの機関との有機的な回路は必須となる。

（2）**ステークホルダーをつなぐネットワーク機能の構築：**d、eおよびhの観点から、fのハブ機能を支援する方策として、例えばAIPネットワークラボやAI R&Dネットワーク（**第4章**の**コラム5**参照）の当初構想の機能的等価物を拡充する。AI研究を担う研究機関・大学などの機関間連携、会議体の設置、プラットフォームの構築などさまざまな形態が考え得るが、このとき、自然科学系の研究者・技術者、産業界、AIガバナンスに関する専門家、セクションを超えた行政担当者、市民・社会・未来の視点との接続など、マルチセクター・マルチステークホルダーが実質的につながる機能が必要である。このネットワーク機能の重要な機能は、各セクターからの知見やアイデア、人材などのプールとして恒常化することである。

（3）**ファンディングプログラムによる加速・拡充：**d、eおよびhの基盤を強化する方策として、例えばRISTEX-HITEのような知見や人材を発掘・育成する機能を持つ研究開発プログラムを、今日の要請に合わせてスケールアップして実施する。このとき、①人文・社会科学系の研究の加速に特化したプログラム化、②基礎・応用研究・先端技術開発プログラムへの融合型、などの実施があり得る。①②ともに、自然科学研究や技術テーマの副次的な位置付けではなく、独立性を持った設計とすることが必要となる。また、個々のプロジェクトレベルの自律的な取り組みを重視しつつ、そこから得られる知見や人材育成などの成果をAIガバナンスやAI研究エコシステムの中に生かしていくためには、プログ

ラムレベルの活動を担う仕組みの担保や、事業・プログラム間や機関間の実質的・継続的な連携の担保が重要である。

今後も「原則も実践も」重要な状況が続くこと、英・米がAI Safety Instituteなどの枠組みを整備し体制を強化しようとしていることを考えると、ガバナンスの上流から下流に至るマルチセクターの取り組みや人材の層を厚くし、かつ相互につながる仕組みが必要と考えられる。さらに日本では、今後数年間のうちに国の科学技術・イノベーション基本計画やAI研究に関わる重要政策が区切りを迎えるタイミングであり、AIガバナンスに関する研究・実践の在り方を再設計するチャンスが到来している。

B.3　AIロボット駆動科学の状況

本節では、研究開発課題③「AI駆動型プロセス革新」における現在の中心的ターゲットであるAIロボット駆動科学の研究事例や政策動向、および、科学研究における基盤モデル・生成AI活用の現状を概観する。

B.3.1　AIロボット駆動科学の取り組み事例

AIロボット駆動科学のグランドチャレンジとして、「2050年までに生理学・医学分野でノーベル賞級の科学的発見をできるAIシステムを作る」ことを目標に掲げた「Nobel Turing Challenge」が2016年に提唱された[67), 68)]。このAIを活用した科学的発見のためのエンジンを作るというグランドチャレンジは国際的な目標になりつつあり、英国のアラン・チューリング研究所では、The Turing AI Scientist Grand Challengeプロジェクトを2021年1月にスタートさせた（Nobel Turing Challenge提唱者の北野もメンバーの一人）。日本国内でも、科学技術振興機構（JST）の未来社会創造事業で本格研究フェーズに移行した「ロボティックバイオロジーによる生命科学の加速」（研究開発代表者：高橋恒一）、ムーンショット型研究開発事業のムーンショット目標3に採択された「人とAIロボットの創造的共進化によるサイエンス開拓」（プロジェクトマネージャー：原田香奈子）などが推進されている（**図B–3–1**）。

生命科学分野では、ロボットによる実験の自動化とAIによる条件探索や最適化を組み合わせた「ロボティックバイオロジー」の取り組みで、科学の自動化に向けた要素技術の開発が進められている。2022年6月には理化学研究所を中心とする研究チームが、汎用ヒト型ロボットLabDroid「まほろ」と新開発の最適化アルゴリズムを組み合わせた自動実験システムを構築し、iPS細胞（人工多能性幹細胞）から網膜色素上皮細胞（RPE細胞）への分化誘導効率を高める培養条件を、人間の介在なしに発見できることを実証した。この系にとどまらず、さまざまな生命科学実験において、これまで人間が試行錯誤をする必要のあった工程が、AIとロボットの導入により自律的に遂行可能になると期待される。材料科学や創薬の分野では、AIによる膨大な可能性からの探索・絞り込みと、ロボティクスによる合成・評価の自動化を組み合わせた「スマートラボ」システムの構築が行われている（**表B–3–1**）。創薬分野ではAstraZenecaのiLab[69)]、材料・化学分野ではトロント大学のSelf-driving Lab[70)]や東京工業大学のデジタルラボラトリー[71)]などが知られている。

AIロボット駆動科学とシミュレーション科学が重なるAI・シミュレーション融合の開発・応用も取り組みが進んでいる。計算コストが高い数値シミュレーションを機械学習モデルで代替（サロゲート）することで高速化し、目的とする新化合物を探索・絞り込む取り組みは「バーチャルスクリーニング」とも呼ばれる。ここで、機械学習モデルの訓練データは数値シミュレーションの結果である。材料科学分野では深層学習モデルを用いたシミュレーションプラットフォームとしてPreferred Computational Chemistry（PFCC：Preferred NetworksとENEOSの共同出資会社）のMatlantisが知られている[72)]。

物理学分野では、深層学習モデル、ニューラルネットワークを物理現象の理解のために用いるという取り組みが活発化し[73), 74), 75)]、物性物理学や重力理論などの分野で成功事例が報告されている。国内では、学術変革領域研究（A）に「学習物理学の創成」（領域代表：橋本幸士）が採択された。ここでは、機械学習と物理学を融合して基礎物理学を変革し、新法則の発見や新物質の開拓につなげることを目指している。また、ボルツマンマシンや拡散モデルなど、物理学のモデルが深層学習の発展につながる流れも見られる。

一方で、人間の認知能力を超えた超多次元の規則性の発見は、もし人間に理解できないとしたら、それを科学として許容して良いのかという議論も起きている[76)]。科学とは何かという基本的な問題や、科学コミュニティーや社会による受容の問題も併せて考えていくことが必要である。

図 B–3–1　　AI ロボット駆動科学の国内関連プロジェクト

表 B–3–1　　AI ロボット駆動科学の具体的事例

名称	主な実施者	内容
iLab	AstraZeneca	Molecular AI とシームレス統合により設計、製造、試験、分析サイクル（DMTA）を高効率化
Self-driving Lab	トロント大学	AI、ロボット工学、HPC を組み合わせて新しい材料や分子の発見を加速
デジタルラボラトリー	東工大	機械学習と定常動作を繰り返す機械を融合した自律的物質探査ロボットシステム
A-Lab[77]	LBNL	ロボット・AI の導入により、材料研究のペースを100倍に加速
AICHEM[78]	USTC	文献読み込み、実験設計、自己最適化する「ロボット化学者」
ASTRAL Lab[79]	Samsung	原料粉末の秤量、混合、焼成、XRD 評価まで自動実験するラボ

B.3.2　政策動向

研究 DX、データ駆動科学、AI・ロボット駆動科学、AI for Science、Automated Research Workflow などのさまざまな概念で表現される科学へのデータ、AI、ロボットの活用を目指す動きは、近年の基盤モデル・生成 AI のブームにより国内外でその動きが活発化している。基盤モデルの科学活用に関しては、既存の調整済みの基盤モデル（ChatGPT など）をそのまま利用する方法、既存の基盤モデルをもとに対象分野のデータを追加学習させてその分野に適応させる方法、独自のコーパスを用いた事前学習を行って独自モデルを開発し利用する方法などが模索されている。現在発表されている多くの事例では、言語モデルに内部表現を獲得させ、その表現を用いて構造予測・物性予測をする仕組みになっている。 Transformer アーキテクチャーは時系列データからの学習にも効果的であり、気象予測や異常検知などのタスクで成果を上げ始めている。

米国科学アカデミー[80]、日本学術会議[81] といった学術界での検討の他、創薬や材料など産業界でも取り組みが進む。 Microsoft Research 社は、英国、中国、オランダに拠点を置く AI4Science チームを 2022 年に設立し、機械学習と物理学・化学・生物学などの世界的な専門家を擁して AI 科学の推進に取り組む。2023 年 12 月には欧州委員会が EU における「AI in Science」の政策の必要性を訴える 58 ページのポリシーブリーフを発行。科学における AI 活用の戦略的重要性と EU における政策的な推進を提言した[82]。

B.3.3 タンパク質言語モデル

AIロボット駆動科学における基盤モデル・生成AIの活用として、特に急速に発展・普及しているのが、タンパク質言語モデル（Protein Language Models）である。言語モデルは、自然言語の文章を構成する各単語の特徴表現と文の生成確率関数を学習する統計モデルであるが、タンパク質言語モデルは各アミノ酸を「単語」とするタンパク質の一次配列という「文章」の特徴表現と文章の生成確率関数を学習する深層学習モデルである。これまでタンパク質言語モデルはword2vecや畳み込みニューラルネットワーク（CNN）などに基づくアーキテクチャーの提案がなされてきたが、近年ではTransformerベースが中心である。

タンパク質の表現学習はこれまでアミノ酸配列の構造を反映したベクトル（記述子）や多重配列アラインメント（Multiple Sequence Alignment: MSA）などが用いられてきたが、タンパク質言語モデルではその特徴表現をデータから獲得する。約2400万配列からなるタンパク質配列データベースUniRefから学習したタンパク質言語モデルUniRepは、その埋め込みベクトルにアミノ酸の物理化学的性質が反映され、分類や変異導入効果の予測などの下流タスクに利用可能なことが示されている[83)]。この表現学習は、アミノ酸配列のマスクされた部分を周辺のアミノ酸残基の情報（自然言語における「文脈」に相当）から推測させるタスクによる自己教師あり学習として行われ、アミノ酸配列中の残基の出現パターンの学習と言える。大規模なTransformerベースのタンパク質言語モデルとしてはESM（Evolutionary Scale Modeling）[84)]、ProtTrans[85)]、ProtT5[86)] などさまざまなものが提案され、獲得された特徴表現は種の分類や変異導入効果の予測などの下流タスクに応用可能である。

表 B–3–2　深層学習とタンパク質立体構造予測

名称	主な開発者	言語モデル	特徴
AlphaFold2	DeepMind	なし	多重配列アライメント（MSA：似た配列の情報、共進化関係を推定可能）やテンプレート構造（似た配列の既知の立体構造）などの特徴量を利用。
ColabFold	MPI、ソウル大、東大など	なし	Google Colaboratory上で高速動作するAF2。 MMSeqs2（ウェブ版）の転用によりMSA取得時間を数分程度まで短縮。
RoseTTAFold	ワシントン大、ハーバード大など	なし	3トラックのネットワークを使用して「1D配列レベル、2D距離マップレベル、3D配位レベルの情報が次々と変換され組み入れられる。
OmegaFold	Helixon	なし	MSAを用いず高速計算が可能。精度はRoseTTAFoldを上回り、AF2と同程度。オーファンタンパク質やMSAがノイジーな抗体も予測。
RGN2	ハーバード大、コロンビア大など	Amino BERT	MSAを用意できないタンパク質（unalgined protein）のアミノ酸配列に潜在している構造情報を学習。平均精度はAF2に劣るが最大で6桁程度高速化。
HelixFold	Baidu	DeBERTa	タンパク質言語モデルの出力ベクトル、アテンションマップを構造予測モジュール（Evoformerベース）に入力。MSA不要のため高速。
ESMFold	Meta AI	ESM2	タンパク質言語モデルの出力ベクトル、アテンションマップを構造予測モジュール（Evoformerベース）に入力。MSA不要のため高速。
EMBER3D	ミュンヘン工科大など	ProtT5	タンパク質言語モデルの出力ベクトル、アテンションマップを構造予測モジュール（RoseTTAFoldベース）に入力。 MSA不要のため高速。

近年ではタンパク質言語モデルがMSAに似た情報を獲得していることを利用し、タンパク質の立体構造予測の高速化を狙う研究が盛んである。タンパク質立体構造予測ではAlphaFold2[87]やRoseTTAFold[88]など深層ニューラルネットを用いた高性能の予測モデルが既に標準的ツールとして利用されているが、いずれも入力としてMSAが必要であり、そこが処理時間上のボトルネックになっている。そこでMSAをタンパク質言語モデル（が保持するMSAに類する情報）で代替するという研究が生まれた（表B–3–2）。

タンパク質言語モデルを利用したタンパク質立体構造予測モデルの代表的な事例としてRGN2[89]、HelixFold[90]、ESMFold[91]、EMBER3D[92]が挙げられる。いずれのモデルもタンパク質言語モデルに基づく表現学習モジュールと構造予測モジュールに分かれた構造を持ち、タンパク質言語モデルから特徴ベクトルとアテンションマップが構造計算モジュールに受け渡される。構造予測モジュールは多くの場合AlphaFold2で用いられたEvoformerというTransformerベースの深層学習モデルが元になっている。大規模タンパク質言語モデルESM2ではモデルサイズが15Bを超えると急激にPerplexityが改善し、立体構造予測の精度も改善するなど基盤モデルの創発性と同様に、タンパク質言語モデルの大規模化にはメリットがあると示唆されている。

B.3.4　AIが変える社会的プロセスとしての科学

AI技術は、科学の知識生産の自動化・自律化を進めるだけでなく、研究者が行う研究活動や、社会の中で営まれる科学の制度・システムにも影響をもたらす（表B–3–3）。大規模言語モデルなどを用いた研究支援ツールや、研究投資（ファンディング）やアウトリーチなど、社会や政策とのインターフェイスにおいてもAI導入が始まる。同時に、生成AIが科学の実践に与える悪影響への懸念も指摘されている。研究の評価やファンディングなど、科学の社会的プロセスの中でのAI活用も盛んに議論・試行されている[29]。近年、世界中で進むオープンサイエンス運動とも相まって、科学におけるAI活用が一段と進展していくと予想される。

表B–3–3　社会的プロセスとしての科学におけるAIの活用の可能性

カテゴリー	AI活用が可能・問題になる場面
研究ワークフローのAI支援	・文献・情報を「探す」「出会う」「読む」「まとめる」「書く」工程のAI支援が進展 ・言語の壁を下げる効果への期待も
論文出版	・査読プロセスへのAI導入
研究評価	・種々の研究評価やグラント審査へのAI導入 ・査読者の選定や、大学研究評価へのAI導入
科学コミュニケーション	・ファクトチェックのためのツールとしてのAI活用 ・平易な記事の自動生成
研究インテグリティー	・再現性・再生可能性の向上のためのAI活用（トラストマーカー有無の自動判定など）
シチズンサイエンス	・人間の集合知とAI活用の関係性は今後の焦点に
科学の科学	・科学計量学がデータ増大と機械学習やネットワーク科学などの手法を取り入れた「Science of Science（科学の科学）」として新段階に発展

29　OECDは2021年に「AIと科学の生産性」というワークショップを実施し、科学のどこにAIを使えるかを検討した。OECD workshop on AI and the productivity of science（2021）

表 B–3–4　研究支援AIツール

サービス名	概要
Jenni Al	論文などのライティング支援ツール。剽窃チェッカー付き。
Humata AI	チャット型オンラインAIサービス。文章要約、質問回答などが可能。
Consensus	質問に対して論文検索を元に科学的コンセンサスの強さとともに回答を表示。
Scite.ai	引用分析プラットフォーム。研究論文の検索を行える機能を実装。
Elicit.org	質問に対して関連論文と要約を出力する論文検索サービス。
Research Rabbit	関心のある論文のコレクションに基づく文献検索、出版アラート機能など。
Semantic Scholar	論文検索サービス。Allen Institute for AIが開発。
Connected Papers	論文検索アプリケーション。関係性がある論文をグラフ表示。

付録C 専門用語解説

AIアライメント（AI Alignment）：人間・社会の価値観にAIを整合（align）させるための技術的・制度的取り組み全般。将来のAIが自律性を高めいわゆる「汎用人工知能（AGI）」や、人間を超えた「超知能（ASI）」となったときに人間を脅かすのではないかという問題意識を持つ論者の間で使われることも多いが、昨今は、より直近のAIシステムがもたらす問題への対応もAIアラインメントの問題として論じられるようになっている。

AIガバナンス（AI Governance）：AIがもたらす倫理的・法的・社会的課題（ELSI）を考慮しつつAI技術によるインパクトを最大化するための、技術・組織・社会システムの設計と運用。本プロポーザルの付録B.2では、AIガバナンスの活動のうち、法規制やガイドラインなどを含むさまざまなルール形成に焦点を当てた。

RLHF（Reinforcement Learning from Human Feedback）：人間のフィードバックを用いた強化学習によって基盤モデルをチューニングする技術。大量データから学習した確率モデルをそのまま用いた生成では、差別的・性的・暴力的など不適切な応答が起こり得る。これを避けるために、応答を人間が評価し、不適切な応答を抑制したり、好ましい応答に修正したりするためのフィードバックを与えてチューニングする。

基盤モデル（Foundation Model）：Stanford Institute for Hu-man-Centered Artificial Intelligenceによって命名され、「大量かつ多様なデータで訓練され多様な下流タスクに適応できるモデル」と定義されている。大量データからさまざまな関係性を事前学習したトランスフォーマー型の超巨大な深層ニューラルネット構造を用いて実現するのが、現在の主流である。基盤モデル（Foundation Model）と呼ばれる他、言語を主としたものは大規模言語モデル（Large Language Model：LLM）と呼ばれることが多い。

生成AI（Generative AI）：ChatGPTに代表されるような、入力された自然言語（プロンプト）に対して応答するAIである。言語の他に画像・映像・音声・動作などマルチモーダルの応答や、言語以外のモーダルを併用した入力も可能になりつつある。まるで人間のような自然な会話や、専門的な知識・能力を備えているかのような応答を返すと大きな注目が集まっている。

ハルシネーション（Hallucination）：生成AIからの応答において、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現象のこと。生成AIでは、大量データから学習した確率モデルに基づいて、言葉をつないでいくため、必ずしも事実に基づかない内容が、自然な文章として生成され得るということが、この問題を生む。

ブリュッセル効果（Brussels Effect）：欧州連合（EU）の規制が、EU域外の国々や企業に影響を与え、これらの国々や企業がEUの規制を自主的に順守する現象のこと。EUのルールをグローバル企業が順守することでデファクト化につながるという側面も見られる。

参考文献

1) McKinsey & Company, “The economic potential of generative AI: The next productivity frontier” (June 14, 2023). https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction
2) 内閣府 総合科学技術・イノベーション会議 第2回基本計画専門調査会（2015年1月22日）資料5「基礎研究力の現状について：基礎研究の位置づけ」https://www8.cao.go.jp/cstp/tyousakai/kihon5/2kai/siryo5-1.pdf
3) Christian Terwiesch, “Would Chat GPT Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course”, Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania (January 17, 2023). https://mackinstitute.wharton.upenn.edu/2023/would-chat-gpt3-get-a-wharton-mba-new-white-paper-by-christian-terwiesch/
4) Tiffany H. Kung, et al., “Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models”, *PLOS Digit Health* 2(2): e0000198 (February 9, 2023). https://doi.org/10.1371/journal.pdig.0000198
5) OpenAI, “GPT-4 Technical Report”, arXiv:2303.08774 (March 15, 2023). https://doi.org/10.48550/arXiv.2303.08774
6) Rishi Bommasani, et al., “On the Opportunities and Risks of Foundation Models”, arXiv:2108.07258 (August 16, 2021). https://doi.org/10.48550/arXiv.2108.07258
7) 科学技術振興機構 研究開発戦略センター,「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報告書 CRDS-FY2023-RR-02（2023年7月）. https://www.jst.go.jp/crds/report/CRDS-FY2023-RR-02.html
8) Wayne Xin Zhao, et al., “A Survey of Large Language Models”, arXiv:2303.18223 (March 31, 2023). https://doi.org/10.48550/arXiv.2303.18223
9) Jingfeng Yang, et al., “Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond”, arXiv:2304.13712 (April 26, 2023). https://doi.org/10.48550/arXiv.2304.13712
10) 森健・林裕之,「日本のChatGPT利用動向（2023年6月時点）～若年層を中心に利用率が高まる～」, 野村総合研究所レポート（2023年6月22日）. https://www.nri.com/jp/knowledge/report/lst/2023/cc/0622_1
11) Jared Kaplan, et al., “Scaling Laws for Neural Language Models”, arXiv:2001.08361 (January 23, 2020). https://doi.org/10.48550/arXiv.2001.08361
12) Jason Wei, et al., “Emergent Abilities of Large Language Models”, arXiv: 2206.07682 (June 15, 2022). https://doi.org/10.48550/arXiv.2206.07682
13) Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo, “Are Emergent Abilities of Large Language Models a Mirage?”, arXiv: 2304.15004 (April 28, 2023). https://doi.org/10.48550/arXiv.2304.15004
14) Long Ouyang, et al., “Training language models to follow instructions with human feedback”, arXiv: 2203.02155 (March 4, 2022). https://doi.org/10.48550/arXiv.2203.02155
15) カテライアメリア・井出和希・岸本充生,「生成AI（Generative AI）の倫理的・法的・社会的課題（ELSI）

論点の概観：2023年3月版」, 大阪大学社会技術共創研究センター ELSI NOTE No.26（2023年4月）. https://doi.org/10.18910/90926
16）岸本充生・カテライアメリア・井出和希,「生成AIの倫理的・法的・社会的課題（ELSI）論点の概観：2023年4～8月版 ーグローバルな政策動向を中心にー」, 大阪大学社会技術共創研究センター ELSI NOTE No.30（2023年9月）. https://doi.org/10.18910/92475
17）新保史生,「EUのAI整合規則提案 ―新たなAI規制戦略の構造・意図とブリュッセル効果の威力」,『情報法制レポート』2022年2巻（2022年2月14日）, pp. 71-81. https://doi.org/10.57332/jilis.2.0_71
18）科学技術振興機構 研究開発戦略センター,「科学技術未来戦略ワークショップ報告書：次世代AIモデルの研究開発 ～技術ブレークスルーとAI×哲学～（2023年11月23日・12月20日開催）」, CRDS報告書 CRDS-FY2023-WR-03（2024年2月）.
19）Stanford University Human-Centered Artificial Intelligence, “Artificial Intelligence Index Report 2023” (April 2023). https://aiindex.stanford.edu/report/
20）科学技術振興機構 研究開発戦略センター,「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－」, CRDS報告書 CRDS-FY2019-SP-08（2020年3月）. https://www.jst.go.jp/crds/report/CRDS-FY2019-SP-08.html
21）科学技術振興機構 研究開発戦略センター,「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～」, CRDS報告書 CRDS-FY2021-SP-03（2021年8月）. https://www.jst.go.jp/crds/report/CRDS-FY2021-SP-03.html
22）科学技術振興機構 研究開発戦略センター,「俯瞰ワークショップ報告書：エージェント技術」, CRDS報告書 CRDS-FY2021-WR-11（2022年3月）. https://www.jst.go.jp/crds/report/CRDS-FY2021-WR-11.html
23）科学技術振興機構 研究開発戦略センター,「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」, CRDS報告書 CRDS-FY2022-SP-03（2022年9月）. https://www.jst.go.jp/crds/report/CRDS-FY2022-SP-03.html
24）石川冬樹・他,『機械学習工学』(講談社, 2022年）.
25）科学技術振興機構 研究開発戦略センター,「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」, CRDS報告書 CRDS-FY2018-SP-03（2018年12月）. https://www.jst.go.jp/crds/report/CRDS-FY2018-SP-03.html
26）丸山文宏,「人とAIの関係性を考える」,『AIデジタル研究』第7号（2023年4月）, pp. 4-13. https://www.ccmc.ac.jp/wp-content/themes/ccmc/pdf/AIDigital07.pdf
27）National Academies of Sciences, Engineering, and Medicine, *Human-AI Teaming: State-of-the-Art and Research Need* (The National Academies Press, 2022). https://doi.org/10.17226/26355
28）三宅陽一郎・清田陽司・大内孝子・他,「レクチャーシリーズ：AI哲学マップ」,『人工知能』(人工知能学会誌）36巻1号（2021年1月）～37巻6号（2022年11月）. https://doi.org/10.11517/jjsai.36.1_74
29）三宅陽一郎・大内孝子・清田陽司,「AI哲学マップ：総論（前編・中編・後編）」,『人工知能』(人工知能学会誌）38巻1号（2023年1月）～28巻3号（2023年5月）. https://doi.org/10.11517/jjsai.38.1_56
30）三宅陽一郎,「人工知能学会 AI哲学マップ総括セッション」, 人工知能学会全国大会第37回（2023年6月6日）. https://speakerdeck.com/miyayou/ren-gong-zhi-neng-xue-hui-aizhe-xue-matupuzong-gua-setusiyon-jiang-yan-zi-liao
31）三宅陽一郎,『人工知能のための哲学塾』(ビー・エヌ・エヌ新社, 2016年）.

32）三宅陽一郎,『人工知能のための哲学塾：東洋哲学篇』(ビー・エヌ・エヌ新社, 2018年）.
33）三宅陽一郎,『人工知能のための哲学塾：未来社会篇 ～響きあう社会、他者、自己～』(ビー・エヌ・エヌ新社, 2020年）.
34）出口康夫,『AI親友論』(徳間書店, 2023年）.
35）鈴木貴之,『人工知能とどうつきあうか：哲学から考える』(勁草書房, 2023年）.
36）鈴木貴之,『人工知能の哲学入門』(勁草書房, 2024年）.
37）Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014). (邦訳：倉骨彰訳,『スーパーインテリジェンス：超絶AIと人類の命運』, 日本経済新聞出版社, 2017年)
38）山川宏・他,「レクチャーシリーズ：シンギュラリティとAI」,『人工知能』(人工知能学会誌)32巻4号(2017年7月）～33巻6号（2018年11月）. https://doi.org/10.11517/jjsai.32.4_590
39）高橋恒一,「将来の機械知性に関するシナリオと分岐点」,『人工知能』(人工知能学会誌)33巻6号(2018年11月）, pp. 867-872. https://doi.org/10.11517/jjsai.33.6_867
40）科学技術振興機構 研究開発戦略センター,「人工知能研究の新潮流 ～日本の勝ち筋～」, CRDS報告書 CRDS-FY2021-RR-01（2021年6月）. https://www.jst.go.jp/crds/report/CRDS-FY2021-RR-01.html
41）The Alan Turing Institute, “Annual Report 2022–23” (2023). https://www.turing.ac.uk/about-us/annual-report-2022-23
42）Yunfan Gao, et al., “Retrieval-Augmented Generation for Large Language Models: A Survey”, arXiv: 2312.10997 (December 18, 2023). https://doi.org/10.48550/arXiv.2312.10997
43）今泉允聡,『深層学習の原理に迫る：数学の挑戦』(岩波書店, 2021年）.
44）Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011). (邦訳：村井章子訳,『ファスト＆スロー：あなたの意思はどのように決まるか?』, 早川書房, 2014年)
45）松尾豊,「知能の2階建てアーキテクチャ」,『認知科学』(日本認知科学会誌）29巻1号（2022年）, pp. 36-46. https://doi.org/10.11225/cs.2021.062
46）Tomoyasu Horikawa and Yukiyasu Kamitani, “Generic decoding of seen and imagined objects using hierarchical visual features”, *Nature Communications* Vol. 8, Article number 15037 (May22, 2017). https://doi.org/10.1038/ncomms15037
47）Satoshi Nishida and Shinji Nishimoto, “Decoding naturalistic experiences from human brain activity via distributed representations of words”, *NeuroImage* Vol. 180, Part A (October 15, 2018), pp. 232-242. https://doi.org/10.1016/j.neuroimage.2017.08.017
48）中原裕之,「社会知性を実現する脳計算システムの解明：人工知能の実現に向けて」,『人工知能』(人工知能学会誌）32巻6号（2017年11月）, pp. 863-872. https://doi.org/10.11517/jjsai.32.6_863
49）田中慎吾・坂上雅道,「推移的推論の脳メカニズム―汎用人工知能の計算理論構築を目指して―」,『人工知能』(人工知能学会誌）32巻6号（2017年11月）, pp. 845-850. https://doi.org/10.11517/jjsai.32.6_845
50）Anthony Zador et al. “Catalyzing next-generation Artificial Intelligence through NeuroAI”, *Nature Communications* Vol.14, Article No.1597 (March 2023). https://doi.org/10.1038/s41467-023-37180-x
51）村田真悟・他,「特集：自由エネルギー原理とAI」,『人工知能』(人工知能学会誌)38巻6号(2023年11月), pp. 777-854. https://doi.org/10.11517/jjsai.38.6_778
52）一藁秀行・他,「特集：予測に基づくロボットの動作学習」,『ロボ學』(日本ロボット学会誌)40巻9号(2022年11月）, pp. 738-806. https://doi.org/10.7210/jrsj.40.738
53）尾形哲也,『ディープラーニングがロボットを変える』(日刊工業新聞社, 2017年）.

54）長井志江,「自閉スペクトラム症の特異な視覚世界を再現する知覚体験シミュレータ」,『精神看護』19巻1号（2016年1月）, pp. 59-63. https://doi.org/10.11477/mf.1689200182
55）西田豊明,『AIが会話できないのはなぜか：コモングラウンドがひらく未来』(晶文社, 2022年）.
56）三宅陽一郎,「大規模 デジタルゲームにおける人工知能の一般的体系と実装 －FINAL FANTASY XVの実例を基に－」,『人工知能学会論文誌』35巻2号（2020年）, pp. B-J64_1-16. https://doi.org/10.1527/tjsai.B-J64
57）Michael Ahn, et al., “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances”, arXiv: 2204.01691 (2022). https://doi.org/10.48550/arXiv.2204.01691
58）Anthony Brohan, et al., “RT-1: Robotics Transformer for Real-World Control at Scale”, arXiv: 2212.06817 (2022). https://doi.org/10.48550/arXiv.2212.06817
59）Yevgen Chebotar and Tianhe Yu, “RT-2: New model translates vision and language into action”, Google DeepMind Blog (July 28, 2023). https://deepmind.google/discover/blog/rt-2-new-model-translates-vision-and-language-into-action/
60）Open X-Embodiment Collaboration, “Open X-Embodiment: Robotic Learning Datasets and RT-X Models” (2023). https://robotics-transformer-x.github.io/
61）立岩陽一郎・楊井人文,『ファクトチェックとは何か』(岩波書店, 2018年）.
62）Xinyi Zhou and Reza Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities”, *ACM Computing Surveys* Vol. 53, No. 5 (September 28, 2020), Article No. 109. https://doi.org/10.1145/3395046
63）笹原和俊,『ディープフェイクの衝撃：AI技術がもたらす破壊と創造』(PHP研究所, 2023年）.
64）Darius Afchar, et al., “MesoNet: a Compact Facial Video Forgery Detection Network”, *Proceedings of IEEE International Workshop on Information Forensics and Security* (WIFS 2018, Hong Kong, December 11-13, 2018). https://doi.org/10.1109/WIFS.2018.8630761
65）Huy H. Nguyen, Junichi Yamagishi and Isao Echizen, “Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos”, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2019, Brighton, May 12-17, 2019). https://doi.org/10.1109/ICASSP.2019.8682602
66）科学技術振興機構 研究開発戦略センター,「戦略プロポーザル：科学技術・イノベーションの土壌づくりとしてのELSI/RRI：戦略的な科学技術ガバナンスの実現に向けて」, CRDS報告書 CRDS-FY2023-SP-01（2023年5月）. https://www.jst.go.jp/crds/report/CRDS-FY2023-SP-01.html
67）北野宏明,「人工知能がノーベル賞を獲る日, そして人類の未来-究極のグランドチャレンジがもたらすもの」,『人工知能』(人工知能学会誌）31巻2号（2016年3月）, pp. 275-286.
68）Hiroaki Kitano, “Nobel Turing Challenge: creating the engine for scientific discovery”, *npj Systems Biology and Applications* Vol. 7, Article No. 29 (June 18, 2021). https://doi.org/10.1038/s41540-021-00189-3
69）Michael Kossenjans, “AstraZeneca iLab: The automated lab of the future”. https://www.astrazeneca.com/r-d/our-technologies/ilab.html
70）Tabassum Siddiqui, “U of T receives $200-million grant to support Acceleration Consortium's ‘self-driving labs’ research”, U of T News (April 28, 2023). https://www.utoronto.ca/news/u-t-receives-200-million-grant-support-acceleration-consortium-s-self-driving-labs-research
71）Ryota Shimizu, et al., “Autonomous materials synthesis by machine learning and robotics”, *APL Materials* Vol. 8, Issue 11 (November 18, 2020). https://doi.org/10.1063/5.0020370
72）So Takamoto, et al., “Towards universal neural network potential for material discovery

applicable to arbitrary combination of 45 elements", *Nature Communications* Vol. 13, Article No. 2991 (May 30, 2022). https://doi.org/10.1038/s41467-022-30687-9
73) 田中章詞・富谷昭夫・橋本幸士 ,『ディープラーニングと物理学：原理がわかる、応用ができる』(講談社, 2019年) .
74) 小林亮太・岡本洋・山川宏 (編) ,「特集：物理学とAI」,『人工知能』(人工知能学会誌) 33巻4号 (2018年7月) , pp. 391-448.
75)「シリーズ：人工知能と物理学」,『日本物理学会誌』74巻1号～11号 (2019年) . https://www.jps.or.jp/books/gakkaishi/seriesai.php
76) Alex Davies, et al., "Advancing mathematics by guiding human intuition with AI", *Nature* Vol. 600 (December 1, 2021), pp. 70-74. https://doi.org/10.1038/s41586-021-04086-x
77) Robert F. Service, "AI-driven robots start hunting for novel materials without help from humans", *Science* Vol. 380, Issue 6642 (April 21, 2023), p. 230. https://doi.org/10.1126/science.adi3613
78) Zhang Tong, "Chinese-made robot is pushing the boundaries of chemical and clean energy research", South China Morning Post (October 14, 2022). https://www.scmp.com/news/china/science/article/3195880/chinese-made-robot-pushing-boundaries-chemical-and-clean-energy
79) Jiadong Chen, et al., "Navigating phase diagram complexity to guide robotic inorganic materials synthesis", arXiv: 2304.00743 (April 3, 2023). https://doi.org/10.48550/arXiv.2304.00743
80) National Academic of Sciences, Engineering, and Medicine, *Automated Research Workflows for Accelerate Discovery: Closing the Knowledge Discovery Loop* (The National Academic Press, 2022). https://doi.org/10.17226/26532
81) 日本学術会議,「研究DXの推進－特にオープンサイエンス、データ利活用推進の視点から－に関する審議について」(2022年12月23日) . https://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-25-k335.pdf
82) European Commission, "AI in Science: Harnessing the power of AI to accelerate discovery and foster innovation", EU Publications (December 14, 2023). https://doi.org/10.2777/401605
83) Ethan C. Alley, et al., "Unified rational protein engineering with sequence-based deep representation learning", *Nature Methods* Vol. 16, No. 12 (October 21, 2019), pp. 1315-1322 (2019). https://doi.org/10.1038/s41592-019-0598-1
84) Alexander Rives, et al., "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences", *Proceedings of the National Academy of Sciences (PNAS)* Vol. 118, No. 15 (April 5, 2021), e2016239118. https://doi.org/10.1073/pnas.2016239118
85) Ahmed Elnaggar, et al., "ProtTrans: Toward understanding the language of life through self-supervised learning", *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 44, No. 10 (October 1, 2022), pp. 7112-7127. https://doi.org/10.1109/TPAMI.2021.3095381
86) Colin Raffel, et al., "Exploring the limits of transfer learning with a unified text-to-text transformer", *The Journal of Machine Learning Research* Vol. 21, No. 1, Article No. 140 (January 1, 2020), pp. 5485-5551. https://doi.org/10.5555/3455716.3455856
87) John Jumper, et al., "Highly accurate protein structure prediction with AlphaFold", *Nature*

Vol. 596, pp. 583-589（July 15, 2021）. https://doi.org/10.1038/s41586-021-03819-2
88）Minkyung Baek, et al., “Accurate prediction of protein structures and interactions using a three-track neural network”, *Science* Vol. 373, Issue 6557（August 19, 2021）, pp. 871-876. https://doi.org/10.1126/science.abj8754
89）Ratul Chowdhury, et al., “Single-sequence protein structure prediction using a language model and deep learning”, *Nature Biotechnology* Vol. 40（October 3, 2022）, pp. 1617-1623. https://doi.org/10.1038/s41587-022-01432-w
90）Xiaomin Fang, et al., “HelixFold-Single: MSA-free Protein Structure Prediction by Using Protein Language Model as an Alternative”, arXiv: 2207.13921（2023）. https://doi.org/10.48550/arXiv.2207.13921
91）Zeming Lin, et al., “Evolutionary-scale prediction of atomic-level protein structure with a language model”, *Science* Vol. 379, Issue 6637（March 16, 2023）, pp. 1123-1130. https://doi.org/10.1126/science.ade2574
92）Konstantin Weissenow, et al., “Ultra-fast protein structure prediction to capture effects of sequence variation in mutation movies”, bioRxiv（November 18, 2022）. https://doi.org/10.1101/2022.11.14.516473
93）Rafael Rafailov, et al., “Direct Preference Optimization: Your Language Model is Secretly a Reward Model”, arXiv: 2305.18290（May 29, 2023）. https://doi.org/10.48550/arXiv.2305.18290
94）Kawin Ethayarajh, et al., “KTO: Model Alignment as Prospect Theoretic Optimization”, arXiv: 2402.01306（February 2, 2024）. https://doi.org/10.48550/arXiv.2402.01306

なお、本プロポーザル中に記載されたURLは、特記のあるものを除いて、2024年2月13日時点のものである。

作成メンバー

総括責任者	木村 康則	上席フェロー	CRDS システム・情報科学技術ユニット
リーダー	福島 俊一	フェロー	CRDS システム・情報科学技術ユニット
メンバー	市川 類	フェロー	CRDS 企画運営室 安全安心グループ
	尾崎 翔	フェロー	CRDS システム・情報科学技術ユニット
	木賀 大介	フェロー	CRDS ライフサイエンス・臨床医学ユニット
	坂本 明史	主任専門員	戦略研究推進部 ICT グループ
	嶋田 義皓	フェロー	CRDS システム・情報科学技術ユニット
	濱田 志穂	フェロー	CRDS 企画運営室 総合知・イノベーショングループ
	丸山 隆一	フェロー	CRDS 企画運営室 横断・融合グループ
	茂木 強	フェロー	CRDS システム・情報科学技術ユニット

戦略プロポーザル **CRDS-FY2023-SP-03**

次世代AIモデルの研究開発

STRATEGIC PROPOSAL

Research and Development on Next Generation AI Models

令和 6 年 3 月　March 2024
ISBN 978-4-88890-894-8

国立研究開発法人科学技術振興機構　研究開発戦略センター
Center for Research and Development Strategy, Japan Science and Technology Agency

〒102-0076 東京都千代田区五番町 7 K's 五番町
電話　03-5214-7481
E-mail crds@jst.go.jp
https://www.jst.go.jp/crds/

本書は著作権法等によって著作権が保護された著作物です。
著作権法で認められた場合を除き、本書の全部又は一部を許可無く複写・複製することを禁じます。
引用を行う際は、必ず出典を記述願います。
なお、本報告書の参考文献としてインターネット上の情報が掲載されている場合には、本報告書の発行日の 1 ヶ月前の日付で入手しているものです。
上記日付以降後の情報の更新は行わないものとします。

This publication is protected by copyright law and international treaties.
No part of this publication may be copied or reproduced in any form or by any means without permission of JST,
except to the extent permitted by applicable law.
Any quotations must be appropriately acknowledged.
If you wish to copy, reproduce, display or otherwise use this publication, please contact crds@jst.go.jp.
Please note that all web references in this report were last checked one month prior to publication.
CRDS is not responsible for any changes in content after this date.

ISBN 978-4-88890-894-8

FOR THE FUTURE OF
SCIENCE AND
SOCIETY

CRDS

https://www.jst.go.jp/crds/