ChatGPT

JST-CRDS「AIエージェントの科学研究への展開」深掘り調査報告書

エグゼクティブサマリー

JST研究開発戦略センター（CRDS）が2026年8月7日に公表したショートレポート「AIエージェントの科学研究への展開」は、AI for Scienceの重心が、予測・解析などの**単機能支援**から、文献探索、仮説生成、計算、実験計画、実験、解析、次の行動選択、報告を結ぶ**研究ワークフロー型システム**へ移行していると論じる。著者は杉村佳織、永野智己であり、個人別の所属は本文に記載されていないが、発行主体は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）である。レポートのDOIは `https://doi.org/10.82643/crds-fy2026-sr-02` である。 [1]

本調査の総合判断は、同レポートの方向性は妥当である一方、現時点で「AI科学者」が一般的な科学研究を自律遂行できると評価するのは早い、というものである。2023〜2026年の一次研究は、狭く定義された課題では実質的な進歩を示している。A-Labは17日間に353実験を行い57候補中36物質を合成し、Robinは文献探索・仮説・人間による実験・AI解析を反復して既存薬リパスジルの候補性を示し、The AI Scientistは機械学習研究の着想から論文作成までを接続した。だが、これらは対象領域、探索空間、装置、評価関数、人間の介入点が強く限定されたシステムである。 [2]

特に重要なのは、**自律性と信頼性は別の軸**だという点である。ScienceAgentBenchでは、実論文に由来する102課題に対し、最良エージェントでも独力解決率は32.4%、専門家知識を与えても34.3%にとどまった。SciAgentArenaも、現行エージェントは明確に構造化されたデータ解析では有効だが、開放的な問題設定、新規洞察、長期的な自己探索では不安定だと報告している。したがって、政策や資金配分は「より長く自律動作するエージェント」だけでなく、独立再実行、証拠連鎖、失敗記録、安全停止、人間の承認、第三者追試を組み込んだ**検証可能な研究システム**を対象とすべきである。 [3]

CRDSレポートが挙げるGoogle Co-Scientist、FutureHouse Robin、Sakana AIのThe AI Scientist、A-Labは、同じ「AIエージェント」という呼称でも実態が異なる。Co-Scientistは主として仮説生成・批判・順位付けを行うマルチエージェント、Robinは文献・仮説・解析をつなぎ人間がウェット実験を担う半自律型、The AI Scientistは計算機内で完結する機械学習研究の自動化、A-Labはロボット、X線回折、能動学習を閉ループ化した自律実験室である。この異質性を無視して単一の「自律度」で並べると、技術成熟度を誤認する。 [4]

日本については、政策の方向性自体は既に整いつつある。文部科学省は2026年3月31日に「AI for Scienceの推進に向けた基本的な戦略方針」を策定し、SPReADによる年間約1,000件規模の裾野拡大と、ARiSEによる科学基盤モデル、AIエージェント、次世代AI駆動ラボへの集中投資を両輪としている。ARiSEの戦略ターゲット型では、マテリアル、バイオ、大型研究施設・装置を対象に、1課題当たり10億〜30億円程度などの大型枠が設定されている。今後の核心は、個別デモへの投資を、共通データ基盤、装置API、評価ベンチマーク、独立追試、研究ソフトウェア人材、安全性評価へどこまで接続できるかにある。 [5]

本報告で明示的に扱う調査項目は次のとおりである。

必須調査項目	掲載箇所
指定ページの全文要約、日付、著者・所属	「指定ページの全文要約と批判的読解」

必須調査項目	掲載箇所
指定ページの全参考文献、原典リンク、主要知見	「引用・参照資料の原典検証」
過去5年中心の代表的学術論文・技術報告10件の比較	「最新研究の横断比較」
分野別応用、技術構成、成果、限界、安全性	「分野別応用、技術構成、限界と安全性」
実験設計、評価指標、再現性、ベンチマーク、データ共有	「研究設計・評価基盤と日本の政策・資金配分」
日本の政策・研究資金配分への示唆	同上
オンライン図表・概念図・Mermaid図	「分野別応用」および「研究設計・評価基盤」

調査対象と方法

調査対象は、2026年8月7日公開のJST-CRDS公式HTMLとPDF、および同ページが明示的に掲げる10件の参考文献である。これに、2021年8月以降を原則的な優先期間としつつ、AIエージェント、マルチエージェント、ツール使用型エージェント、自律実験室、科学エージェント評価に関する原著論文、査読論文、公式技術報告、政府資料を追加した。情報は2026年8月7日までに公開されたものに限定し、企業発表については査読済み科学的証拠と区別して評価した。 [6]

評価軸は、単なる「自動化率」ではなく、研究工程の範囲、人間の介入、自律的な次行動選択、外部ツール・装置への接続、成果の科学的妥当性、独立再現性、証拠追跡性、安全制御、第三者評価の有無とした。この区分は、CRDS自身が組織ごとに情報収集、課題設定、仮説生成、実験計画、実行、条件調整、解析、次アクション、報告、知識蓄積への対応が異なると整理していることに基づく。 [7]

証拠の強さは概ね、査読済み原著かつ物理実験・外部評価を伴うもの、査読済みだが計算機内評価に限られるもの、プレプリントまたは開発組織による技術報告、企業の運用発表の順に区別した。ただし、査読済みであっても、開発者自身による少数課題評価、LLMによる自動評価、単一研究室での検証は、一般化可能性を保証しない。Co-Scientistの15課題に対するElo評価も、論文自身が「独立した真値に基づく評価ではない」と明記している。 [8]

指定ページの全文要約と批判的読解

書誌情報。 指定ページの正式名称は、ショートレポート AI for Science Spotlights Vol.2「AIエージェントの科学研究への展開」。公開日は2026年8月7日、著者は杉村佳織・永野智己である。個人ごとの所属肩書は本文・署名欄に記載されておらず、機関としての発行主体はJST-CRDSである。公式HTMLは `https://www.jst.go.jp/crds/column/kaisetsu/shortreport2.html`、PDFは `https://www.jst.go.jp/crds/column/kaisetsu/pdf/CRDS-FY2026-SR-02.pdf`、DOIは `https://doi.org/10.82643/crds-fy2026-sr-02` である。 [9]

冒頭の主要メッセージ。 レポートは三つのポイントを提示する。第一に、科学向けAIは研究者の問いに答える支援ツールから、文献調査、仮説生成、計算、データ解析、実験計画、結果評価を連続処理するシステムへ発展している。第二に、その具体例としてCo-Scientist、The AI Scientist、FutureHouseの科学エージェント、A-Labを挙げる。第三に、今後は自律性の向上そのものよりも、科学的妥当性、再現性、判断過程の追跡可能性、人間の承認、安全性を備えたシステム設計が課題になるとする。 [1]

AIエージェントの定義。 本文は科学研究向けAIエージェントを、研究者が示した目的から必要な作業を組み立て、文献データベース、研究データ、専門ソフトウェア、計算環境を利用して一連の処理を進めるAIシステ

ムと定義する。単一モデルだけでなく、文献探索、コード生成、解析、条件提案などを担う複数エージェントと、それらを調整する監督エージェントからなるマルチエージェント型も含む。発展の背景には基盤モデルの推論・コーディング能力と、データベース、ソフトウェア、計算資源をAIから呼び出す技術の進歩があると説明する。[10]

表による異質性の整理。 レポートの表1は、Google、FutureHouse、ローレンス・バークレー国立研究所、Sakana AI、三井化学、Matlantis、東京大学、パナソニック インダストリーを、10の研究工程に分けて比較している。GoogleとFutureHouseは情報収集・仮説・次アクションに強く、LBNLのA-Labは実行・条件調整・解析に強い。三井化学は情報収集・報告、Matlantisは実験計画に相当する計算設定・実行・解析、東京大学とパナソニックは物理実験の実行を中心とする。表には「未確認」も含まれ、各組織が公表した機能の整理であって、共通ベンチマークによる性能比較ではない。[7]

海外事例。 Co-Scientistは、研究者の目標と既存証拠から仮説を作成し、複数エージェントが生成、批判、順位付け、改良を反復する。RobinはCrow、Falcon、Finchなどを連携させ、仮説、実験計画、データ解析、追加仮説を反復する。A-Labは計算材料データ、文献から抽出した合成知識、機械学習、能動学習、ロボット実験、X線回折を統合し、失敗結果に応じて次の条件を変更する。この最後の点を、固定手順を繰り返す「自動化」と結果に応じて次行動を変える「自律化」の差として位置付けている。[11]

国内事例。 Sakana AIのThe AI Scientistは、機械学習分野でアイデア生成、文献探索、コード、計算実験、解析、図表、論文、査読を一体化する。三井化学は化学構造式の画像と文献テキストを統合して化合物情報を調査し、初期検証では従来約1カ月の調査を約1日に短縮したとする。MatlantisはClaude CodeやCodexを原子レベルシミュレーターに接続し、構造緩和、分子動力学、反応経路、結晶構造予測などのSkillsを与える。東京大学等は薄膜合成、X線回折、解析を接続し、パナソニック インダストリーは24時間365日運転可能な自動実験設備を整備している。[12]

将来像。 レポートは、人間、複数AI、計算機、装置が分担・並行して研究を進めることで、探索可能な仮説や条件が増えると予測する。同時に、AI生成物が増えれば人間の検証能力がボトルネックになり、コード再実行、異手法による解析、物理的追試、モデル・データ・コード・条件・判断履歴の保存が必要になると指摘する。高リスク実験、外部システムへのアクセス、成果公開には人間の承認を組み込み、AIを科学者の代替ではなく、科学者が扱える探索範囲を拡張する研究基盤として設計すべきだと結論する。[10]

批判的評価。 この結論は、最新研究の証拠と整合する。一方、レポートは短報であり、体系的レビューの検索式、採否基準、性能指標、証拠グレードを提示していない。参考文献にはNature原著、研究機関の公式発表、企業ニュース、採用サイトの記事が並列され、科学的検証の強度が異なる。また、表1は機能の有無を示すが、成功率、誤り率、人間工数、費用、エネルギー、独立再現率、安全違反率を比較していない。したがって、優れた技術動向マップではあるが、各システムの優劣や一般的自律性を実証する評価報告ではない。[13]

さらに、「エンドツーエンド」という表現には注意が必要である。Robinでは物理実験は人間が実施し、The AI Scientistでは最終原稿自体は人間が修正しなかったものの、提出候補の選別、コードが動くか、テーマに合うか、書式が正しいかの確認は人間が行った。したがって、現状の代表例は「人間不在の科学者」ではなく、**人間が目標、境界、装置、候補選別、ウェット作業、承認を担う条件付き自律システム**と捉えるのが正確である。[14]

引用・参照資料の原典検証

指定ページが掲げる参考文献は以下の10件である。原典へのURLと、原典に基づく主要知見、証拠上の留意点を併記する。

CRDS番号	原典とURL	原典の主要知見	検証上の留意点
[1]	Gottweis et al., “Accelerating scientific discovery with Co-Scientist,” Nature (2026). https://doi.org/10.1038/s41586-026-10644-y	Gemini系モデルを基盤とする監督・専門エージェントが仮説生成、批判、順位付け、改良を反復。生命科学の複数事例で人間研究者が候補仮説を実験検証した。 15	15課題のElo評価は独立真値ではなく自動評価。著者自身も厳格な査読、外部DBによる事実確認、引用精度向上が必要とする。 16
[2]	Lu et al., “Towards end-to-end automation of AI research,” Nature 651, 914–919 (2026). https://doi.org/10.1038/s41586-026-10265-5	アイデア、文献、コード、計算、解析、論文、査読を接続。3本をICLR 2025ワークショップへ提出し、1本が平均6.33点で採択水準を上回った。 17	ワークショップ採択率は70%で、3本中2本は基準未達。内部評価では3本ともICLR本会議水準に達せず、引用誤り、実装ミス、浅いアイデア等が確認された。 18
[3]	Ghareeb et al., “A multi-agent system for automating scientific discovery,” Nature 655, 497–505 (2026). https://doi.org/10.1038/s41586-026-10652-y	Robinが文献探索、候補生成、実験提案、解析を反復。リパスジルは対照比1.89倍のRPE貪食能増加を示し、一次ヒトRPE細胞でも候補性を確認した。 19	主にin vitro証拠で、論文自身がin vivo検証を必要とする。物理実験は人間が担当した。
[4]	FutureHouse, “Demonstrating end-to-end scientific discovery with Robin.” https://www.futurehouse.org/research/demonstrating-end-to-end-scientific-discovery-with-robin-a-multi-agent-system	Crow、Falcon、Finchを連携し、研究構想から投稿まで約2.5カ月。コードと例示軌跡を公開した。 20	開発組織自身による説明であり、「知的枠組み全体がAI駆動」という表現は自己評価。査読原著[3]と併読すべきである。
[5]	Sakana AI「世界初、100%AI生成の論文が査読通過」 https://sakana.ai/ai-scientist-first-publication-jp/	AI Scientist-v2が仮説、コード、実験、解析、原稿を生成し、1本が査読者平均6.33点となった。倫理審査と会議運営者の許可を得て実施された。 21	事前合意により撤回され、メタレビューは未実施。開発元自身が、開示規範や査読負荷の問題を指摘している。

CRDS番号	原典とURL	原典の主要知見	検証上の留意点
[6]	Szymanski et al., “An autonomous laboratory for the accelerated synthesis of inorganic materials,” Nature 624, 86–91 (2023). 原著：https://doi.org/10.1038/s41586-023-06734-w、訂正：https://doi.org/10.1038/s41586-025-09992-y	A-Labは計算材料データ、文献モデル、ロボット、XRD、能動学習を統合し、17日間、353実験、57標的中36件の合成に成功した。[22]	CRDS参考文献欄は原著の書誌に対して訂正記事のDOIを付している。原著の正しいDOIは2023年のもの。原著には2026年1月の著者訂正がある。[23]
[7]	三井化学「研究開発の文献調査を革新する生成AIエージェントを開発」https://jp.mitsuichemicals.com/jp/release/2026/2026_0302/index.htm	構造式画像とテキストを統合し、化合物名、用途、物性、製造法、条件を調査。初期検証で調査時間80%以上削減、約1カ月から約1日への短縮を報告した。[24]	社内実証の企業発表であり、課題数、正解率、誤抽出率、外部評価などは公開ページからは十分確認できない。
[8]	Matlantis「AIエージェント連携機能を提供開始」https://matlantis.com/ja/news/release-260527/	Claude Code、Codexと原子レベルシミュレーターを接続し、自然言語からコード生成。Skillsは構造緩和、MD、反応経路、結晶予測等を提供する。[25]	専門ツールへの自然言語インターフェースとして有力だが、自律的仮説検証や発見成功率の独立評価ではない。
[9]	東京大学「欲しい物質を自動的・自律的に合成する」https://www.s.u-tokyo.ac.jp/ja/press/10741/	モジュール型測定システム、標準データ形式、クラウド、AI解析を統合したデジタルラボを構築。関連論文はDigital Discovery、DOI https://doi.org/10.1039/D4DD00326H。[26]	CRDSは薄膜合成、XRD、解析、$LiCoO_2$条件最適化を紹介する。個別装置の相互運用性は強みだが、汎用科学エージェントとは区別が必要。[10]
[10]	パナソニック インダストリー「自動実験で切り拓く材料開発の未来」https://recruit.industry.panasonic.com/story/1478/	門真拠点のスマートラボが24時間365日の自動実験を可能にし、反復作業を削減すると説明する。[27]	採用サイト上のプロジェクト紹介であり、査読論文、探索性能、再現率、自律的条件変更の詳細は示されていない。

原典検証から、CRDSの参考文献構成には二つの系列があることが分かる。第一はCo-Scientist、The AI Scientist、Robin、A-Labのような査読済み研究で、第二は三井化学、Matlantis、パナソニック等の導入・製品・組織事例である。後者は研究現場への実装動向を把握するには重要だが、学術的性能を示す一次実験とは証拠レベルが異なる。CRDS短報を政策判断に用いる場合は、「研究成果」と「導入発表」を別レイヤーで読む必要がある。[28]

A-Labの書誌上の不整合は、AIエージェント研究で重視すべき証拠追跡性を象徴する。CRDS PDFに記載された `10.1038/s41586-025-09992-y` は2026年の訂正記事で、2023年原著そのもののDOIは `10.1038/s41586-023-06734-w` である。これはレポートの技術的主張を無効にするものではないが、機械可読な引用検証と原典照合を必須化すべきことを示す。[29]

最新研究の横断比較

以下は、CRDSの主張を検証・補強するために選んだ2023〜2026年の代表的研究・技術報告10件である。研究工程の広さだけでなく、実証方法、成功指標、限界を比較した。

研究・年	分野とエージェント構成	主な成果	主要な限界・評価
Co-Scientist, Nature, 2026	生命科学。監督エージェントと生成、批判、順位付け、進化等の専門エージェント。文献検索・外部ツール・研究者フィードバックを利用。	薬剤再目的化、線維症標的、薬剤耐性遺伝子移動などで仮説を提示し、一部を人間が実験検証。[15]	15課題評価は小規模で、Eloは自動評価かつ独立真値ではない。一般科学への外挿は未検証。[8]
Robin, Nature, 2026	生物医学。Crow/Falconによる文献・候補探索、Finchによるデータ解析、ワークフロー調整。物理実験は人間。	dAMD候補探索でリパスジル、KL001、ABCA1関連機序を提示。リパスジルはRPE細胞の貪食能を対照比1.89倍に増加。[19]	in vitro中心。in vivo、臨床有効性、別研究室での再現は未確立。検索エージェントを除くと架空引用が増えた。
The AI Scientist, Nature, 2026	機械学習。アイデア生成、Semantic Scholar検索、並列ツリー型コード探索、計算、論文、AI査読。	3本中1本がICLRワークショップ査読で平均6.33点。研究工程の計算機内接続を実証。[18]	ワークショップ採択率70%。3本とも本会議水準未達。引用誤り、実装不整合、浅い着想、幻覚が残る。
Science One Framework, 技術報告、2026	計算研究。文献グラフ、並列探索、読み取り専用実験記録、主張検証器、Chain-of-Evidence Audit。	開発者報告では基準系の架空引用率が最大21%だったのに対し、同システムは架空引用ゼロ、スコア再実行一致を達成。[30]	プレプリント・実験プロトタイプで、主にシステム最適化・ML課題。独立評価と科学分野横断検証が必要。

研究・年	分野とエージェント構成	主な成果	主要な限界・評価
A-Lab, Nature, 2023	無機材料。DFT、文献抽出モデル、ロボット合成、XRD、ML解析、能動学習。	17日連続、353実験、57標的中36件、成功率63%。 22	空気安定材料、利用可能前駆体、温度等は人間が限定。副生成物、計算誤差、前駆体揮発などで失敗。
Coscientist, Nature, 2023	化学。GPT-4、ウェブ・文書検索、コード実行、クラウドラボ、ロボットAPIを複数LLMで統括。	既知化合物の合成計画、装置文書利用、液体ハンドラー制御、Pd触媒反応最適化など6課題を実証。 31	課題と装置が事前設定され、危険化学、高価な装置、例外処理に対する一般的安全性は未確立。
ChemCrow, Nature Machine Intelligence, 2024	化学。GPT-4に18の専門ツール、反応予測、逆合成、物性計算、ロボット実行を接続。	防虫剤と3種の有機触媒の合成を計画・実行し、新規発色団探索を支援。 32	ツールAPIの品質・可用性に依存。専門家評価の規模、危険反応の防止、非専門家による誤用が課題。
BioPlanner, EMNLP, 2023	生物実験計画。自然言語プロトコルを擬似コード化し、BioProtで計画能力を評価。	検索した擬似コードを用いて新規プロトコルを生成し、実験室で一例を成功裏に実行。 33	長期・多段階計画は依然難しく、単一成功例は広範な実験自動化を意味しない。
ScienceAgentBench, ICLR 2025	データ駆動科学。44査読論文から102課題、4分野。コード、実行、コストを評価。	最良エージェントは3試行で独力32.4%、専門家知識付き34.3%。o1-previewは42.2%だがコストは他モデルの10倍超。 34	現行エージェントがデータ分析コード生成でも多数失敗し、汎用エンドツーエンド科学には距離があることを示す。
SciAgentArena, プレプリント、2026	単一細胞・空間オミクス、創薬、遺伝学等。約200課題、対話環境、段階別検証。	構造化され評価基準が明確な解析では有効性を確認。 35	新規洞察、自己主導探索、開放的研究問題では性能が不均一。プレプリント段階。

横断比較から得られる第一の結論は、成功事例の多くが、**目的関数を明示でき、実行結果を短時間で機械評価できる領域**に集中していることである。機械学習ベンチマーク、物性計算、XRD解析、細胞アッセイ、反応収率などは、エージェントにフィードバックを返しやすい。反対に、研究課題そのものの重要性、概念的飛躍、理論的説明、長期的因果推論、社会的価値判断は定量化が難しく、現状の実証は限定的である。 36

第二に、マルチエージェント化それ自体が科学的信頼性を保証するわけではない。役割分担、討論、自己批判は候補数や探索幅を増やすが、複数エージェントが同じ基盤モデル、学習データ、検索源を共有すれば、誤りや偏りも相関する。Co-Scientistの自動Elo評価やRobinのLLM judgeは有用な工程内評価である一方、独立した専門家、別モデル、別コード、別研究室による検証を代替しない。 37

第三に、最先端は「文章を上手く生成するAI」から「証拠を保存し、主張と実行物を結び付けるAI」へ移りつつある。Science OneのChain-of-Evidenceは、論文中の数値を再実行結果に、手法記述を実コードに、引用を実在文献に結び付ける。これはCRDSが求める判断履歴・追跡可能性を、抽象的原則から機械評価可能な設計要件へ変える試みである。ただし、現状は開発元の技術報告であり、第三者による再評価が不可欠である。 30

分野別応用、技術構成、限界と安全性

分野	代表事例と技術構成	実証された成果	技術的限界	倫理・安全性課題
生命科学・創薬	Co-Scientist：仮説生成・批判・順位付け。Robin：文献エージェント＋解析エージェント＋人間ウェットラボ。	仮説候補の実験検証、既存薬再目的化候補、追加実験の反復提案。 38	培養系から動物・臨床への外挿、データ品質、因果機序、患者多様性、独立再現。	病原体操作、デュアルユース、患者データ、誤った治療仮説、過度な自動化。AI scientistの誤操作がバイオセーフティ上の危険を生み得る。 39
化学・合成	ChemCrow/Coscientist：LLM＋反応DB＋逆合成＋計算＋ロボットAPI。	合成計画、反応実行、条件最適化、装置文書の自律利用。 40	装置差、試薬在庫、スケールアップ、異常系、API・DB依存。	発熱、爆発、有毒物質、危険な生成物、非専門家による実行。人間規制・エージェント整合・環境側制御の三層防御が必要。 41
無機・材料科学	A-Lab：DFT＋文献モデル＋ロボット＋XRD＋能動学習。東京大学：モジュール装置＋標準データ。Matlantis：自然言語＋原子シミュレーター。	連続自律合成、失敗結果による条件更新、自然言語から専門計算。 42	探索範囲を人間が限定。相同定、純度、組成、装置較正、異種装置接続がボトルネック。	高温・高圧・有害前駆体、装置衝突、誤った相同定、企業データ・知財流出。
機械学習・計算科学	The AI Scientist：アイデア、コード、計算、論文。Science One：証拠グラフ、並列探索、主張検証。	計算機内で研究工程を接続し、査読・再実行まで機械化可能であることを実証。 43	ベンチマーク過適合、評価器攻略、コードと論文の不一致、引用幻覚、計算費用。	査読システムの飽和、AI生成開示、著者責任、不正なスコア最適化、再現不能論文の大量生成。
文献・知識統合	三井化学：構造式画像＋文献テキスト＋化学DB＋ウェブ検索＋報告。	社内初期検証で調査時間80%以上削減、約1カ月から約1日。 24	正解率、見逃し率、構造認識誤り、非公開文献、古いDB、同義語処理の外部評価が不足。	機密情報の外部送信、ライセンス違反、誤引用、誤った先行技術判断。

分野	代表事例と技術構成	実証された成果	技術的限界	倫理・安全性課題
自動実験・産業ラボ	パナソニックのスマートラボ等：装置統合、連続稼働、反復作業自動化。	24時間365日の運転と人的反復作業の削減。 27	「連続自動運転」と「結果に基づく自律的研究判断」は別。科学的成果指標が未公表。	労働安全、保守責任、非常停止、サイバー侵入、ログ改ざん、装置ベンダーロックイン。

安全性は、アクセス権を一度与えた後にAIへ注意を促すだけでは確保できない。Nature Communicationsのリスク分析は、①ユーザー意図、②化学・生物・放射線・物理・情報などの科学領域、③人間・自然環境・社会経済への影響という三軸で危険を評価し、「人間による規制」「エージェント自身の危険認識・整合」「環境・装置側の規制」からなる三層の保護を提唱する。したがって、モデルの安全プロンプトだけでなく、装置側の物理インターロック、試薬・温度・圧力の許容範囲、権限分離、サンドボックス、監査ログ、緊急停止が必要である。 44

科学的方法上のリスクも物理安全と同等に重要である。MesseriとCrockettは、AI導入により研究量が増えても、人間が実際以上に理解したと錯覚し、特定の手法、問い、データ、観点が支配する「科学的モノカルチャー」が生じ得ると論じる。同じ基盤モデルを多数の研究者が使う場合、見かけ上は独立した仮説でも、学習データや検索ランキングに由来する相関した偏りを持つ可能性がある。したがって評価には、精度だけでなく、代替仮説の多様性、引用源の集中度、少数派理論の探索率も含めるべきである。 45

参考となるオンライン図表は以下である。

図表	参照URL	読み取るべき点
CRDS表1「研究プロセス対応」	`https://www.jst.go.jp/crds/column/kaisetsu/pdf/CRDS-FY2026-SR-02.pdf#page=2`	同じAIエージェントでも対応工程が大きく異なる。 7
Co-Scientist概念図	`https://www.nature.com/articles/s41586-026-10644-y/figures/1`	監督エージェントと専門エージェントの非同期協調。 15
The AI Scientistワークフロー	`https://www.nature.com/articles/s41586-026-10265-5/figures/1`	アイデア、ツリー探索、原稿、査読の計算機内閉ループ。 46
Robin概念図	`https://www.nature.com/articles/s41586-026-10652-y/figures/1`	文献・仮説・人間実験・解析の分業構造。 47
A-Lab概念図	`https://www.nature.com/articles/s41586-023-06734-w/figures/1`	計算、文献、ロボット、XRD、能動学習の物理閉ループ。 48
ChemCrow概念図	`https://www.nature.com/articles/s42256-024-00832-8/figures/1`	LLMが専門ツールをThought–Action–Observationで利用する構成。 32
Science One / Chain-of-Evidence	`https://research.google/blog/science-one-framework-a-verifiable-autonomous-research-framework-via-chain-of-evidence/`	主張、引用、コード、実行結果を結ぶ証拠連鎖と監査。 30

研究設計・評価基盤と日本の政策・資金配分

CRDSレポートの問題提起を実証可能な研究課題へ変換するには、エージェントの「賢さ」を単一スコアで測るのではなく、研究工程、成果、再現性、効率、安全性、認識論的多様性を分離して評価する必要がある。ScienceAgentBenchが示す低い完遂率、Science Oneが示す引用・コード・数値の不整合、AI Scientistが報告する実装ミスと引用幻覚は、最終論文の文章品質だけを評価してはならないことを示している。 [49]

推奨する実験設計。 評価実験は、まず研究課題、成功条件、禁止行為、利用可能ツール、計算・試薬予算、停止条件を事前登録する。課題は、既知答えのある再現課題、将来時点の予測課題、未公開の新規課題に分け、学習データ汚染を避けるため公開時点で時間分割する。比較群は、①人間のみ、②基盤モデルのみ、③ツール使用型単一エージェント、④マルチエージェント、⑤人間とエージェントの協働とし、計算量、時間、情報アクセス、実験予算を可能な限り揃える。

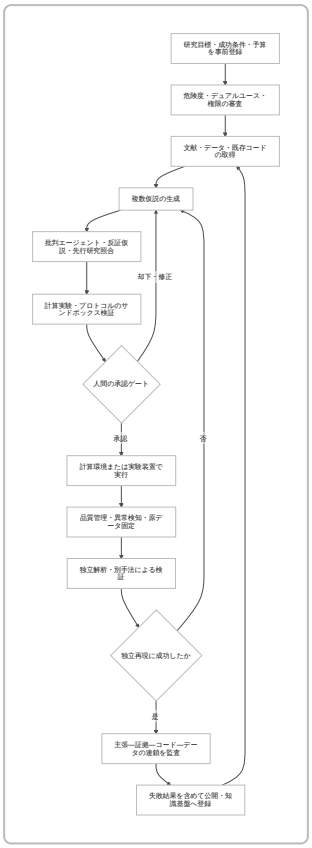
各条件は複数モデル、複数乱数シード、複数研究チームで反復し、評価者をシステム名から盲検化する。ウェット実験では、開発研究室と独立研究室の二段階追試を行い、肯定結果だけでなく、失敗、無効結果、停止、危険提案、修正履歴を公開する。AIが候補を多数生成する場合は、多重検定、逐次選択バイアス、選択的報告を補正しなければならない。

推奨評価指標。

評価領域	推奨指標
科学的妥当性	事前登録された仮説の支持率、効果量、信頼区間、予測校正、Brier score、偽陽性率、専門家間一致度
新規性・重要性	時点固定の先行研究検索後に残る新規性、代替仮説数、情報利得、追試価値、後続研究への寄与
再現性	コード再実行成功率、数値一致率、別環境での再現率、独立研究室での追試率、データ・依存関係完全率
証拠追跡性	主張―証拠リンク率、引用実在率、手法―コード一致率、モデル・プロンプト・ツール・データ履歴の記録率
効率	壁時計時間、人間作業時間、計算費、試薬費、エネルギー、目標到達までの実験数、サンプル効率
自律性	人間介入回数、承認待ち時間、AIが選択した次行動の割合、例外からの回復率、停止判断の適切性
安全性	危険提案率、権限違反率、禁止ツール呼び出し率、物理インターロック作動数、ニアミス、ロールバック成功率
多様性・頑健性	仮説・手法・引用源の多様性、モデル交換時の結果安定性、分布外課題性能、反証仮説の探索率

CRDSが求める「記録・共有」は、単に最終コードをGitHubに置くことでは足りない。最低限、基盤モデル名と版、システムプロンプト、エージェント構成、ツール版、検索時刻、取得文献、入力・出力、乱数、実行環境、失敗ログ、人間の介入、装置較正、原データ、解析コード、判断理由を機械可読形式で保存する必要がある。モデルが更新されるサービス型AIでは、同じ名称でも挙動が変わるため、API版、日時、温度・推論設定、応答ハッシュを保存すべきである。Science Oneの証拠連鎖は、この要件を論文作成時ではなく主張生成時に組み込む考え方として参考になる。 [30]

推奨する統制付き研究フローは次のとおりである。



このフローでは、AIは仮説候補の数を増やし、計算と反復を高速化するが、危険性審査、装置実行、結果公開の前に独立した承認ゲートを置く。低リスクの計算タスクでは承認を簡略化し、高リスクの化学・生物・大型装置では多者承認、最小権限、物理インターロックを必須にする「リスク比例型ガバナンス」が適切である。 41

日本の政策環境。 文部科学省は2026年3月31日に正式な「AI for Scienceの推進に向けた基本的な戦略方針」を策定した。同方針はAI for Scienceを単なる効率化ではなく研究構造の転換と位置付け、研究者の創造性、共通研究インフラ、分野横断・国際連携、倫理・信頼性・透明性を柱とする。JST-CRDSの「AI for Scienceの動向2026」も、AIが研究プロセス全体に浸透する「第5の科学パラダイム」の可能性を整理している。 50

裾野拡大型のSPReADは年間約1,000件の研究利用を想定し、AI初心者も含めた伴走支援とコミュニティ形成を掲げる。一方、ARiSEは科学基盤モデル、AIエージェント、次世代AI駆動ラボを対象とするフラグシップ型で、ドメイン研究者とAI・数学・システム研究者の共同代表体制を求める。ARiSEはマテリアル、バイオ、大型研究施設・装置を戦略ターゲットとし、国際ベンチマーク、AI-readyデータ、実験装置統合、Human–AI Co-scientistを事業要件に含めている。 51

予算上限は、マテリアル関連で30億円程度を各1課題、別の30億円程度を1課題、10億円程度を4課題、バイオで20億円程度を3〜4課題、大型施設・装置で20億円程度を1〜2課題とされる。単純合計では、戦略ターゲット型の提示枠だけで約180億〜220億円規模になり得る。これは日本がAI駆動研究を小規模実証ではなく研究基盤として構築しようとしていることを示す。 52

ただし、大型投資が個別の基盤モデル、ロボット、デモ設備に偏ると、研究機関間の再利用性が低くなる。日本では装置メーカー、材料産業、大型研究施設、高品質実験データが比較優位になり得るため、競争力の源泉は汎用LLMの規模だけではなく、**装置を安全に呼び出せる標準API、校正済みデータ、失敗データ、プロ**

ベナンス、遠隔実験基盤に置くべきである。ARiSEも、マテリアルの国際ベンチマーク、継続的データ蓄積、バイオのAI-readyデータ、大型施設の自動自律化を明示している。 53

政策実装に向け、本報告は次の試案的な資金ポートフォリオを提案する。これは政府の公式配分ではなく、検証可能性を重視した分析上の提案である。

配分の目安	投資対象	理由
45%	マテリアル、生命科学、大型施設等の分野別フラグシップ	国際的に意味のある科学成果と実装を創出する中核
25%	NII RDC、HPCI、SINET、装置API、データ標準、遠隔ラボ等の共通基盤	個別課題終了後も再利用できる公共財を形成
12%	ベンチマーク、独立追試、第三者監査、失敗結果リポジトリ	開発チームによる自己評価への偏りを減らす
8%	化学・生物安全、サイバーセキュリティ、レッドチーム、社会科学・倫理	高自律化に伴う物理・情報・社会リスクを管理
10%	Research Software Engineer、ラボ自動化技術者、データキュレーター、保守	モデル開発後の統合、運用、再現性を持続させる

特に、各大型課題の研究費の一部を「独立検証勘定」として開発チームから切り離し、別機関によるコード再実行、装置追試、安全性評価、引用監査に充てることが望ましい。採択時KPIも、論文数や処理時間削減だけでなく、独立再現率、データ・コード完全率、安全違反ゼロ、装置・モデル交換可能性、外部研究者の再利用数を含めるべきである。これはARiSEが掲げる国際ベンチマークと、CRDSが強調する再現性・追跡可能性を実効化する。 54

結論と推奨

JST-CRDSのショートレポートは、AI for Scienceを「個々の予測モデル」の問題から、「人間、複数AI、データベース、計算資源、実験装置をどう組み合わせて研究を進めるか」という研究システム設計の問題へ正しく移している。その最大の価値は、AIエージェントを単一技術として扱わず、仮説生成型、ワークフロー統合型、ツール使用型、自律実験型が並行発展していることを示した点にある。 4

一方、原典を精査すると、現在の成果は「限定された科学工程での高性能」を示すものであり、「一般的な科学者の代替」を示すものではない。A-Labは物理閉ループで強いが探索空間は人間が定め、Robinは知的工程を広く担うがウェット実験は人間が行い、The AI Scientistは計算機内で工程を完結できるが本会議水準の研究を安定して生み出せない。ScienceAgentBenchとSciAgentArenaは、日常的な科学コード、開放的問題、長期探索に大きな未解決領域があることを示す。 55

したがって、今後の第一優先課題は、自律度競争ではなく**検証可能性競争**である。全ての主張を文献、原データ、コード、実行ログ、装置条件へ結び、第三者が再実行できるシステムを評価上優遇すべきである。架空引用率、手法とコードの不一致、再実行不能な数値は、論文の文章品質とは別に機械監査しなければならない。 30

第二に、研究成果の評価単位を「AIが論文を書いたか」から「人間とAIの組み合わせが、同じ資源でより良い科学を生んだか」へ改めるべきである。評価には、正しさ、新規性、独立再現、探索効率、安全性、人間工数、費用、エネルギー、仮説多様性を含める。査読通過は一指標にすぎず、特に採択率の高いワークショップでの一例を、広範な科学的信頼性の証明として扱ってはならない。 18

第三に、日本は、巨大汎用モデルの開発競争だけを追うのではなく、材料、生命科学、計測、大型施設、製造装置という既存資産を、AIが安全かつ再現可能に利用できる研究基盤へ変換するべきである。装置API、標準試料、校正データ、失敗データ、NII RDC、SINET、HPCI、遠隔実験、Research Software Engineerを横断的に整備することが、個々のプロジェクトを超えた国際的不可欠性につながる。文部科学省の戦略とARiSEはこの方向を既に含むため、次の課題は共通仕様と持続的運用への具体化である。 56

第四に、独立追試と安全性を研究費の周辺業務ではなく、主要成果物として資金化すべきである。大型課題には外部再現チーム、監査チーム、化学・生物安全チームを配置し、開発チームとは別の予算・指揮系統を持たせる必要がある。高リスク工程では、モデルの指示にかかわらず環境側が危険操作を拒否できる物理的・ソフトウェア的制約を必須とする。 41

最終的に、AIエージェントが科学にもたらす価値は、人間を工程から排除した割合では測れない。価値は、研究者がより広い仮説空間を探索し、失敗を早期に発見し、証拠を完全に残し、第三者が結果を再現し、社会が安全に利用できるかで決まる。CRDSレポートの「科学者を直ちに置き換えるのではなく、科学者が扱える範囲を広げる研究基盤として普及させる」という結論は、原典と最新ベンチマークを踏まえても、現時点で最も妥当な政策・研究開発原則である。 57

---

1 4 6 9 10 11 12 13 54 57 ショートレポート　AI for Science Spotlights Vol.2「AIエージェントの科学研究への展開」｜CRDSフェローが解説！最新のサイエンス｜特集・短報・コラム｜研究開発戦略センター（CRDS）
https://www.jst.go.jp/crds/column/kaisetsu/shortreport2.html

2 22 23 29 36 42 48 55 An autonomous laboratory for the accelerated synthesis of inorganic materials | Nature
https://www.nature.com/articles/s41586-023-06734-w

3 34 49 [2410.05080] ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery
https://arxiv.org/abs/2410.05080

5 50 AI for Scienceの推進に向けた基本的な戦略方針の策定について：文部科学省
https://www.mext.go.jp/b_menu/shingi/chousa/shinkou/077/aifors_strategy.html

7 28 jst.go.jp
https://www.jst.go.jp/crds/column/kaisetsu/pdf/CRDS-FY2026-SR-02.pdf

8 15 16 37 38 Accelerating scientific discovery with Co-Scientist | Nature
https://www.nature.com/articles/s41586-026-10644-y

14 20 Demonstrating end-to-end scientific discovery with Robin | FutureHouse
https://www.futurehouse.org/research/demonstrating-end-to-end-scientific-discovery-with-robin-a-multi-agent-system

17 18 43 46 Towards end-to-end automation of AI research | Nature
https://www.nature.com/articles/s41586-026-10265-5

19 47 A multi-agent system for automating scientific discovery | Nature
https://www.nature.com/articles/s41586-026-10652-y

21 世界初、100%AI生成の論文が査読通過 「AIサイエンティスト」が達成
https://sakana.ai/ai-scientist-first-publication-jp/

24 三井化学、研究開発の文献調査を革新する生成AIエージェントを開発 | ニュースリリース | 三井化学株式会社
https://jp.mitsuichemicals.com/jp/release/2026/2026_0302/index.htm

25 Matlantis、AIエージェント連携機能を提供開始
https://matlantis.com/ja/news/release-260527/

26 Press Releases - 東京大学 大学院理学系研究科・理学部
https://www.s.u-tokyo.ac.jp/ja/press/10741/

27 自動実験で切り拓く材料開発の未来――衆知を集め未知の技術に挑む「想い」 | パナソニック インダストリー採用サイト
https://recruit.industry.panasonic.com/story/1478/

30 Science One Framework: A verifiable autonomous research framework via Chain-of-Evidence
https://research.google/blog/science-one-framework-a-verifiable-autonomous-research-framework-via-chain-of-evidence/

31 Autonomous chemical research with large language models | Nature
https://www.nature.com/articles/s41586-023-06792-0

32 40 Augmenting large language models with chemistry tools | Nature Machine Intelligence
https://www.nature.com/articles/s42256-024-00832-8

33 [2310.10632] BioPlanner: Automatic Evaluation of LLMs on Protocol Planning in Biology
https://arxiv.org/abs/2310.10632

35 Benchmarking AI Agents for Addressing Scientific Challenges Across Scales
https://arxiv.org/abs/2606.12736?utm_source=chatgpt.com

39 41 44 Risks of AI scientists: prioritizing safeguarding over autonomy | Nature Communications
https://www.nature.com/articles/s41467-025-63913-1

45 Artificial intelligence and illusions of understanding in scientific research | Nature
https://www.nature.com/articles/s41586-024-07146-0

51 SPReAD 1000 - 研究の可能性を、AIで解き放つ｜文部科学省
https://www.mext.go.jp/aifors_spread/

52 53 mext.go.jp
https://www.mext.go.jp/content/20260416-mxt_jyohoka01-000049067_001.pdf

56 AI for Scienceの推進に向けた 基本的な戦略方針
https://www.mext.go.jp/content/20260515-mxt_kibanken01-000049769_8.pdf?utm_source=chatgpt.com