ChatGPT

「仮説を作るAI」は科学者を不要にするのか――科学的発見の自動化と「知的分業」の再設計

エグゼクティブサマリー

野口悠紀雄氏の記事の核心は、AIによって「答え」だけでなく「仮説」まで大量供給されるようになれば、科学者の希少な能力は「新しいアイデアを出すこと」から、「何を研究し、どの仮説を検証し、限られた資源をどこに投じるかを判断すること」へ移る、というものだ。記事はさらに、その最上位には「なぜその問題を研究するのか」「どのような社会を望むのか」という価値判断があり、そこに人間の役割が残ると論じる。[1]

この方向性はかなり正しい。しかし、科学者の役割を「価値判断」にまで後退させて考える必要はない。
2026年時点の最先端システムを見ると、GoogleのCo-Scientistは文献探索、仮説生成、批判、ランキング、改善まで行い、一部の生物医学仮説は実験でも支持された。しかしNature論文自身が、このシステムを明確に「scientist-in-the-loop」と位置づけ、研究目標・制約・候補選択・ウェットラボ検証には専門家を組み込んでいる。著者らも、幻覚、不完全な文献、負の結果へのアクセス不足、研究方向の均質化、再現性危機の悪化を主要なリスクとして認めている。[2]

したがって、現在起きていることをより正確に表現すれば、**「科学者の消滅」ではなく「科学におけるボトルネックの移動」**である。AIは、文献探索、組み合わせ的な発想、候補生成、コード作成、予備解析といった「探索コスト」を劇的に下げつつある。一方で相対的に希少になるのは、問題設定、測定設計、因果識別、対照実験、異常値の解釈、候補の優先順位づけ、安全性判断、独立再現、外的妥当性、研究資源配分、そして最終的な説明責任である。Co-Scientistの実験でも、専門家がウェットラボ実験を優先順位づけしており、in vitroで成功してもin vivoや臨床成功を保証しないと明記されている。[3]

もう一つ重要なのは、**「仮説を作れる」と「科学的発見が成立した」は同義ではない**ことである。
HypoBenchでは、既存のLLM仮説生成法は有効で新規なパターンを見つけられる一方、難易度が上がると、最良の組み合わせでも人工データの正解仮説の38.8%しか回収できなかった。AI Scientist-v2は完全自動で研究を行い、3本中1本がICLRワークショップの平均採択基準を上回ったが、開発者自身の人間による本会議水準の内部評価では3本とも基準未達だった。別研究では、自律AI科学者に不適切なベンチマーク選択、データリーク、指標誤用、事後選択バイアスが確認され、完成論文だけでなくコードと全実行ログの監査が必要だと指摘されている。[4]

本稿の結論は次のようになる。

> **AI時代に科学者の中心職能は、「仮説を思いつく人」から「信頼できる知識生成システムを設計・統治する人」へ拡張される。**
> 発想そのものの希少性は低下する可能性が高い。しかし、どの仮説が本当に新しいのか、何なら反証可能なのか、どの測定が信用できるのか、何を独立検証すべきか、社会として何を優先するかという仕事は、むしろ重要になる。

そして制度設計上は、AIを「科学者の代替者」として評価するより、**AIを含む研究チーム全体の再現性、真正性、監査可能性、探索効率、安全性を評価する方が有益**である。Natureの現在の編集方針も、AIを支援技術と位置づけ、「学術的判断と説明責任は人間から移転できない」と明示している。[5]

記事の要約と批判的読解

記事は、おおむね四段階の論理で構成されている。第一に、AIが既知の情報への回答だけでなく未知の科学問題にも到達し始めたことを「質的転換」とみなす。具体例として、遺伝子に関する発見と約80年間未解決だった数学問題を挙げる。後者については、Natureも2026年5月に、OpenAIのAIがPaul Erdősに関係する約80年前の数学問題を解いた事例を報じている。 [6]

第二に、AIが回答だけでなく仮説そのものを生成・選別できるようになると、それまで期待されていた「人間が問いを作り、AIが答える」という分業も崩れると論じる。GoogleのCo-Scientistを念頭に、AIが「問い→仮説→検討→新しい問い」というサイクルを回し得るとする。実際、Googleが2026年に一般向けに説明しているHypothesis Generation機能も、研究者と研究課題を定義した後、複数AIエージェントが仮説を生成・議論・評価する「アイデア・トーナメント」を行う設計である。 [7]

第三に、記事は「問い」に階層があるとする。「どの遺伝子が病気を起こすか」という対象レベルの問いより上に、「なぜこの病気を研究するか」「なぜそこへ資金と人材を投入するか」、さらに「どのような社会を望むか」という問いがあり、後二者には価値判断が入り込む。したがってAIが目的に沿って最適化できても、「目的そのものを何にするか」は別問題だ、とする。 [1]

最後に、「仮説が不足する世界」から「仮説が過剰供給される世界」へ移れば、希少資源は発想ではなく、**選択・検証・資源配分**になると結論づける。科学者は「新しいアイデアを思いつく人」から「何を検証する価値があるか判断する人」へ変わるべきだ、というのが記事の最も重要な主張である。 [1]

記事の主要主張	暗黙の前提	本稿の評価
AIは未知の答え・仮説を作り始めた	現在の実例が一般的な科学能力へ拡張可能	**概ね正しいが要限定。** 生物医学・機械学習・数学では有力事例があるが、分野横断的な自律性は未証明。 [8]
「人間が問い、AIが答える」分業は崩れる	AIが問いと仮説を反復生成できる	**既に部分的に成立。** Co-Scientist等は反復生成する。ただし研究目標は依然として人間が与える。 [9]
仮説は過剰供給される	生成コストが検証コストより急速に下がる	**非常に重要な洞察。** ただし問題は仮説の「数」ではなく、偽陽性と検証負債も同時に大量生産されること。 [10]
科学者は選別者になる	発想の希少価値が低下する	**方向は正しいが狭すぎる。** 測定、実験設計、因果推論、再現、解釈、安全管理も希少になる。 [3]
最終的な価値判断は人間に残る	AIには価値設定能力がない／乏しい	**能力論より正統性・責任論として捉えるべき。** AIが価値に関する議論を生成できても、誰の価値を採用し、その結果に誰が責任を負うかは社会制度上の問題である。 [11]

記事の修辞にも特徴がある。タイトルの「科学者たちは不要に!?」はまず「職業消滅」という強い脅威フレームを置き、本文後半で「そうではない」と反転させる構造になっている。「数年前には簡単な数学でも間違えたAI」から「80年来の難問を解くAI」への対比は、技術進歩の速度を印象づける。さらに「工場労働者・単純事務→知的労働者」という従来の自動化物語を科学者へ拡張し、読者自身の職業不安と接続する。 [1]

最も巧みなのは「希少性の逆転」である。AIによって回答と仮説が豊富になれば、「何に注意を向けるか」が希少資源になるという議論は、経済学者らしい整理であり、現在のAI科学研究ともよく整合する。逆に弱い点は、「科学者たちは強い危機感を抱いているだろう」といった部分が実証ではなく推測であること、また科学研究を「問い→仮説→答え」にやや単純化し、測定装置、試料品質、対照群、因果識別、研究室の暗

黙知、再現実験などを十分扱っていないことである。記事で紹介された技術論文自体は、むしろこれらの人間・実験系を依然として不可欠な構成要素としている。[12]

したがって記事の命題は、

「問いの価値が上がる」

から一段進めて、

「検証可能な問いへ翻訳し、信頼できる証拠へ変換し、その証拠に基づいて責任ある判断をする能力の価値が上がる」

と修正するのが、2026年時点の技術状況にはより適切である。

「仮説生成AI」はどこまで科学をできるのか

まず重要なのは、「AIが科学的発見をした」という言葉を分解することだ。最低でも次のような証拠段階を区別しなければならない。

証拠段階	何ができれば到達か	現在のAIの状況
文献再結合	既存知識からもっともらしい候補を生成	かなり強い。Co-ScientistやMOOSE-Chemが典型。[13]
隠された既知結果の再発見	学習時点以降の論文等を伏せて既存発見を再構成	化学やAMRで成功例。ただし「未知の自然法則の発見」とは異なる。[14]
計算上の新規結果	コード、シミュレーション、数学的検証等で新しい候補を得る	機械学習・数学では急速に進展。AI Scientist系が典型。[15]
実験的支持	実際の試料・細胞・装置で予測を検証	Co-Scientistで選択された例にin vitro検証がある。[16]
独立再現・外的妥当性	別研究室、別データ、別条件でも再現	まだ限られる。Co-Scientist論文も検証を「preliminary」とする。[17]
臨床・社会的な知識	患者集団・現実環境で有効、安全、一般化可能	現状の仮説生成AIからは大きな距離。in vitro成功も臨床成功を保証しない。[18]

この区別をすると、「AIは科学者を代替した」という主張がなぜ早すぎるかが分かる。**現在最も劇的に自動化されているのは「探索空間を広く速く走査すること」であり、「自然界について信頼できる知識を確立すること」全体ではない。**

主要システムを比較すると以下のようになる。

システム／研究	技術方式	何を自動化するか	注目すべき実績	主な限界・リスク
Schmidt & Lipson	**記号的・進化的探索**	実験データから数式的法則を探索	2009年に自然法則の自動抽出を提示	物理的事前構造が実質的に入っているとの後続批判もあり、「無前提の発見」は注意が必要。 19
Robot Scientist Adam	**記号推論＋実験自動化**	仮説生成→実験→解釈	酵母機能ゲノミクスで2009年に閉ループ科学を実証	対象領域・実験系が強く構造化されている。仮説生成AIそのものはLLM以前から存在した。 20
The AI Scientist	**LLMエージェント＋コード実行**	発想、コード、実験、図、論文、模擬査読	ML三分野でend-to-end研究を自動化	主として計算機上のML研究で、初代は人間作成テンプレートへの依存が大きい。 21
AI Scientist-v2	**LLM＋agentic tree search**	仮説→実験→論文をほぼ一貫自動化	3本をICLRワークショップへ送り、1本が平均採択基準を超過	ワークショップ水準。人間が3本を選び提出。開発者自身の本会議水準評価では全3本が基準未達。引用ミス等も確認。 22
MOOSE-Chem	**LLMマルチエージェント＋文献検索**	背景→着想文献→化学仮説	2024年の51本の高水準化学論文を用い、学習カットオフ後の仮説を「再発見」	評価の中心は既存論文の仮説への類似度。「本当に前例のない未知」の検証とは違う。 23
HypoBench	**ベンチマーク**	仮説生成手法を194データセットで比較	実世界7タスク＋合成5タスク	難化すると最良でも正解仮説回収38.8%。現状の探索能力の不完全さを示す。 24
Google Co-Scientist	**LLMマルチエージェント＋検索・ツール**	生成、批判、ランキング、進化、実験案	AML、肝線維症、AMRで選択仮説を実験評価。一部成功	人間が研究目標、制約、候補選択、実験優先順位を担当。幻覚、文献偏り、負の結果欠落、均質化リスク。 25
Kosmos	**LLMエージェント＋世界モデル＋コード＋文献**	長時間のデータ分析・文献探索・仮説反復	最大12時間、200ロールアウト。著者報告では独立科学者が報告文中の記述79.4%を正確と評価	約8割という数字は高い一方、科学的主張には無視できない誤り余地がある。「6か月相当」は共同研究者の自己評価で、標準化された生産性指標ではない。 26
AI Sleep Co-Scientist	**専門家誘導型マルチエージェント**	仮説、信号処理、統計解析	約12.4万件・50TB超の睡眠データを複数ケースで分析	人間が専門エージェントを指揮し中間出力をレビューする設計。2026年7月時点でプレプリント。 27

Co-Scientistは特に重要である。Geminiを基盤とするGeneration、Reflection、Ranking、Evolution、Proximity、Meta-reviewなどのエージェントが、仮説を「生成→討論→順位づけ→進化」させる。研究者は自然言語で研究目標、制約、望ましい性質を与え、探索方向を途中で変更できる。つまり、AIは単なるチャットボットではなく、**巨大な「仮説探索エンジン」**へ近づいている。 9

一方、その性能評価には注意が必要である。Co-Scientistの専門家評価は11研究課題という小規模なもので、平均noveltyは5点中3.64、impactは3.09だったが、論文自身がこれは客観的ground truthではなく専門家の主観評価だと明記している。さらに大規模な実験検証はコスト上不可能だったため、ウェットラボ候補は専門家が優先順位づけした。 28

実験成功にも階層がある。たとえばAMLの候補薬・組み合わせでin vitroの効果が得られても、Nature論文は、患者の異質性、生物学的複雑性、薬物動態、オフターゲット作用などによって、in vivo有効性や臨床成功は保証されないと強調している。 18

また、「新発見」に見えるものが、実はAIにとって見えなかった既知結果の**再発見**である場合もある。Co-ScientistのAMR例は、AIが2日ほどで到達した仮説が、同時期に別の研究チームが実験的に見つけ、当時まだ査読前だった機構と一致したという強い結果だが、これは「AIが自然界で唯一最初に発見した」とは異なる。むしろ重要なのは、**独立した探索経路が同じ機構へ収束したことがAIの科学的有用性を示す**という点である。 18

さらに、自律化を進めるほど監査上の問題が大きくなる。LuoらはAI Scientist-v2等を調べ、不適切なベンチマーク選択、データリーク、指標の誤使用、結果を見た後の選択バイアスという四つの典型的な失敗モードを検出した。そして、最終論文だけ読むより、AIの**全実行ログとコード**を調べる方が問題発見に有効だと報告している。 29

ここから一つ重要な原則が導ける。

> **AIによる仮説の生産速度が100倍になっても、検証能力が100倍にならなければ、「知識」が100倍になるのではなく、「検証待ち仮説」が100倍になる。**

これは野口氏の「仮説過剰供給」論をさらに進めたものである。AI科学時代の最大の問題は、**アイデア不足ではなく「検証負債（validation debt）」**になる可能性がある。

科学技術の歴史から見える「仕事消滅」ではなく「ボトルネック移動」

「機械が科学者の中核作業へ入ってくる」という現象自体は、2026年に突然始まったわけではない。2009年にはRobot Scientist “Adam” が機能ゲノミクスで仮説生成と実験を自動化し、同じScience誌号ではSchmidtとLipsonが実験データから数式的自然法則を探索するシステムを報告していた。したがって現在の革新は「機械が初めて仮説を作った」ことではなく、**汎用LLM、大規模文献検索、コード実行、マルチエージェント、ロボット実験が一つの柔軟な研究ループへ統合され始めたこと**にある。 30

科学技術史を圧縮すると、次のような流れになる。

時期	技術・出来事	自動化・拡張されたもの	人間側で相対的に重要になったもの
二十世紀後半	計算機・統計ソフト	数値計算、反復計算、大規模統計処理	モデル化、データ設計、解釈、アルゴリズム設計

時期	技術・出来事	自動化・拡張されたもの	人間側で相対的に重要になったもの
1990–2003	Human Genome Project	DNA配列決定の工業的スケール化、大規模データ生成	コンソーシアム運営、データ基盤、ゲノム解析・バイオインフォマティクス。HGPは国際的大規模協働として2003年に完了した。 31
2009	Adam／symbolic discovery	仮説生成、実験、数式法則探索の一部	問題領域の構造化、意味解釈、装置・実験系の構築。 32
2021–	AlphaFold	タンパク質立体構造予測	構造予測そのものから、機能仮説、相互作用、実験設計へ研究時間を再配分。AlphaFold DBはその後2億超の予測構造を提供する規模になった。 33
2024–25	AI Scientist、MOOSE-Chem	アイデア、コード、計算実験、執筆、化学仮説再発見	評価設計、ベンチマーク真正性、選択、監査。 34
2026	Co-Scientistなど	文献→仮説→批判→ランキング→実験案	研究目標、候補選択、実験、検証、再現、倫理・資源配分。 35

AlphaFoldは特に良い比較対象である。タンパク質構造予測の重要部分を機械が劇的に高速化しても、生物学者や構造生物学者そのものが論理的に不要になるわけではない。むしろ予測構造が大量供給されることで、「この構造からどの機構を疑うか」「どの変異を作るか」「どの相互作用を測るか」「予測が壊れる条件は何か」という後段の問いが増える。AlphaFoldのNature論文自体にも入力情報が乏しい場合などの限界があり、予測と実験的知識は同一ではない。 36

Human Genome Projectにも同じ構造が見える。ゲノム配列を大量に読めるようになったことは、「ゲノム研究の終了」ではなく、データ解釈、比較ゲノム、機能ゲノム、統計遺伝学など、さらに巨大な研究領域を生んだ。HGPは20か国規模の共同事業として進められ、2003年に完了した。 31

この歴史から得られる一般則は、

技術がある科学作業の限界費用を下げると、その作業そのものの価値は相対的に下がるが、その出力を入力として使う「次の工程」の価値が上がる

ということである。

AI仮説生成でも同じことが起きる可能性が高い。

従来：

文献を読む → 数個の仮説を考える → 実験を選ぶ

AI時代：

数千の文献を統合 → 数百の候補を生成 → 矛盾・反例を探索 → 数十候補を機械的に予備検証 → 数個だけ高価な実験へ

となる。Co-Scientistはすでにこの後者の方向へ進んでいる。[35]

ただし、ここには危険なフィードバックもある。AIが過去の文献をもとに仮説を生成し、そのAI仮説に基づく論文が大量出版され、次世代AIがそれを学習するようになれば、誤り・出版バイアス・流行が自己増幅する可能性がある。Co-Scientist論文も、オープンアクセス文献への依存、負の実験結果の欠落、矛盾した文献から誤った知見が伝播する可能性、研究方向の均質化を明示的な限界として挙げている。[17]

したがって「科学者が減るか」という人数論より、**科学者の時間配分がどこへ移るか**を見る方が有益である。本稿の予測では、最も圧縮されやすいのは文献サーベイ、定型解析、初期ブレーンストーミング、標準コード、定型的な文章化であり、相対的価値が上がるのは、独創的な測定系、原因と相関の切り分け、例外への感度、現場の暗黙知、高品質データ生成、独立再現、研究ポートフォリオ設計、他者が信頼できる証拠へ変換する仕事である。これは現在のAI科学システムが最も人間へ依存している箇所とも一致する。[37]

AIと科学者の「知的分業」はどう設計すべきか

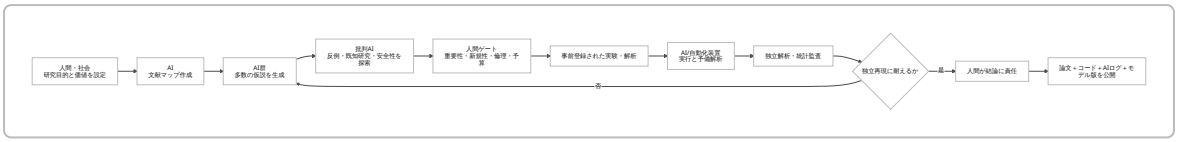
ここで重要なのは、AIと人間の境界を固定的に「AI＝答え、人間＝問い」と決めないことである。性能が向上すれば境界は移動する。代わりに、**どの決定にどの水準の証拠と説明責任が必要か**で分業すべきである。

考えられる主要モデルを比較すると次のようになる。

分業モデル	AIの役割	人間の役割	適する研究	最大の利点	主なリスク
コパイロット型	文献整理、発想、コード、批判、草案	問題・仮説・実験・結論を主導	高リスク臨床研究、探索初期	責任境界が明確	人間の既存バイアスをAIが補強する可能性
仮説ポートフォリオ型	数百候補生成、反例探索、ランキング	評価関数、重要度、安全性、最終候補選択	創薬、材料、ゲノム	AIの探索幅を最大活用	ランキング指標へのGoodhart化、流行の均質化
ゲート付き閉ループ型	仮説→ロボット実験→解析→再仮説	安全・予算・段階移行ごとに承認	標準化可能な化学・材料・細胞実験	高速反復	誤った代理指標を高速最適化する危険
提案者対監査者型	AI-Aが仮説、AI-Bが反証・再解析	人間が対立する証拠を裁定	統計・計算科学、バイオインフォ	自己確認バイアスを抑えやすい	両AIが同じモデル・文献バイアスを共有
社会的ミッション型	与えられた公共目標内で探索・最適化	患者、市民、科学者、政策当局が優先順位を決定	公衆衛生、気候、希少疾患	「何を研究するか」を民主的に統治	合意形成が遅く、価値対立は残る

現在のCo-Scientistに最も近いのは二番目であり、AI Sleep Co-Scientistは一番目と二番目の中間である。Google自身もCo-Scientistを「scientist-in-the-loop」とし、研究者が研究目標、制約、初期案、フィードバックを与える設計としている。[38]

本稿が研究機関の標準モデルとして推奨するのは、**「仮説ポートフォリオ＋提案者対監査者＋人間ゲート」**の組み合わせである。



この方式では、人間はAIの全出力を一つ一つ読む必要がない。重要なのは**ゲートをどこに置くか**である。

第一のゲートは「研究する価値があるか」。ここには疾患負担、科学的重要性、公平性、費用、患者ニーズなどが入る。第二は「実験する価値があるか」で、反証可能性、既存研究との差、事前確率、安全性、費用対情報利得を見る。第三は「知識として主張してよいか」で、統計、独立再現、代替説明、外的妥当性を確認する。この三つは単なるAI能力の問題ではなく、説明責任の配置でもある。

特に重要なのが「ランキングAIを信じすぎない」ことである。仮説生成AIが同じモデルで自分の仮説を評価すると、共通の誤りが自己増幅する可能性がある。Co-ScientistはReflectionやトーナメントでこれを軽減しているが、論文自身が幻覚やバイアスを未解決問題として残している。AI Scientist研究では、評価指標の誤用や事後選択も現実に観察されている。したがって、可能なら生成器と評価器を異なるモデル・異なる情報源・異なる統計手法にし、最終検証は生成過程から独立させるべきである。 39

数学・プログラム検証・シミュレーションのように答えを機械的に検査しやすい領域では、AIへ渡せる範囲は広くなる。一方、生物医学では「細胞で効いた」から「動物でも効く」「人でも効く」「長期的に安全」「医療制度上価値がある」までに複数の飛躍がある。Co-ScientistのAML研究がその違いを明瞭に示している。 18

したがって**自律性はAIモデルの賢さだけで決めるべきではない**。

$$\text{許容できるAI自律性} \propto \frac{\text{結果の機械検証可能性} \times \text{可逆性}}{\text{失敗時の被害} \times \text{検証困難性}}$$

という発想が有用である。

コード改善や定理候補なら自律性を高くできる。一方、患者治療や病原体操作のような領域では、同程度に優秀なAIであっても人間ゲートを厚くするべきである。これは「AIを信頼するか」という抽象的論争より、はるかに運用可能な考え方である。

研究機関・政策・教育・出版制度をどう変えるか

制度上の最大の課題は、現在の科学が「人間が比較的ゆっくり仮説と論文を生産する」前提でできていることだ。AIが数百・数千の仮説、解析、原稿を低コストで生成できるようになると、査読、再現、研究費審査といった人間側の仕組みがボトルネックになる。AI Scientist-v2が一つのシステムから論文をend-to-end生成できることは、この圧力がすでに概念上のものではないことを示している。 40

研究者には、「AIに良い答えを書かせる技術」よりも「AIが間違いやすい実験を設計する技術」が重要になる。つまりAIが支持する仮説を確認するだけでなく、それが誤りなら最も早く壊れる実験、強い代替仮説を区別できる実験を優先すべきである。候補生成時と最終評価時のデータを分離し、AIがテストセットを見た後に研究計画を変更することを防ぐ必要もある。これはAI scientistで実際に確認されたデータリークや事後選択バイアスへの直接的な対策になる。 29

大学・研究所は、AI利用を一律禁止するのではなく、研究リスク別の運用にするべきである。公開論文の要約と、未公表患者データ・企業秘密・研究費申請書の解析ではリスクが全く異なる。日本政府も、2025年12月のAI関連技術に関する指針に加え、2026年3月改定のAI事業者ガイドライン、AISIの安全性・データ品質関

連文書、厚労省の医療データAI利用ガイドラインなどを並行して位置づけている。研究機関には、これらを「研究AI利用規程」へ翻訳した実務ルールが必要である。 41

具体的には、機密データを扱える承認済み環境、入力・出力・モデルバージョン・ツール呼び出しのログ保存、AI利用箇所の申告、人間責任者の署名、バイオセーフティやデュアルユース案件の追加審査を標準化するのが望ましい。これは単に不正防止のためではない。モデル更新で同じプロンプトから違う結果が生まれる以上、将来「AI支援研究を再現する」にはモデル・設定・検索時点の記録が必要になる。AI科学者の失敗を最終論文だけから見抜きにくいという研究結果も、ログ保存の必要性を強く支持する。 29

研究費配分機関は、AIによって仮説生成が安価になるほど、「新規アイデアの数」を評価する制度から「情報価値の高い検証」を評価する制度へ移るべきである。特に、独立再現、負の結果、ベンチマーク構築、高品質データセット、測定装置、AI評価基盤への予算を増やす意味が大きくなる。Co-Scientist論文が負の実験結果へのアクセス不足を主要な限界としていることは示唆的であり、**AI時代にはnegative resultsがこれまで以上に公共財になる**。 17

審査機密性にも注意が必要である。NIHは、研究費申請書の内容をオンライン生成AIへ入力することが秘密保持義務に反するとして、査読者による生成AIを使った申請書分析・査読文生成を禁止している。したがって研究助成機関がAI査読補助を導入する場合も、公開型LLMへ申請書を投入するのではなく、隔離された承認済みシステムを使うことが最低条件になる。 42

学術誌・学会については、AI利用の有無より「何をAIへ委任したか」を記録する方が重要になる。Nature Portfolioの2026年時点の方針は、AI利用をリスク別に扱い、文章改善などの補助利用と、解釈・評価に影響する利用を区別し、学術的判断と説明責任を人間からAIへ移せないとしている。また、AIが生成した仮説・分析・結論を人間由来であるかのように提示することや、査読判断そのものをLLMへ委任することを認めていない。 5

今後さらに一歩進めるなら、AIが研究の中核に関与した論文には「AI Methods Statement」を設け、

使用モデル／バージョン、プロンプトまたは研究目標、アクセス可能だった文献、生成候補数、選択基準、途中で人間が介入した箇所、実験・解析コード、AI実行ログ、失敗した候補、独立検証方法

を記録するのが合理的である。Google自身もHypothesis Generation利用時には、仮説・実験設計・優先順位づけへのAI利用を方法欄等で開示することを推奨している。 43

教育も大きく変える必要がある。知識の暗記が不要になるという意味ではない。AIの誤りを見抜くには「何があり得ないか」を判断できる強い専門知識が要る。むしろ教育で比重を高めるべきなのは、因果推論、実験計画、測定理論、統計的検定、再現性、データ生成過程、AI監査、研究倫理、科学哲学、異分野の概念を翻訳する能力である。Co-Scientistが大量の文献を統合できても、誤った・矛盾した文献まで取り込むという限界は、専門知識の不要化ではなく、**専門知識の役割が「記憶」から「検証」に移る**ことを示している。 17

教育上は、学生に「AIを使うな」ではなく、

> AIに仮説を10個作らせる
> → 最も魅力的な3個を選ばせる
> → その3個を自分で反証する
> → AIに反対論を作らせる
> → 一番安く区別できる実験を設計する
> → 結果を見ずに解析計画を固定する

という演習の方が、AI時代の科学者訓練として有効だろう。

利害関係者ごとの優先策をまとめると、次のようになる。

主体	直ちに始めるべきこと	避けるべきこと
研究者	AI仮説をポートフォリオとして扱い、反証実験・事前登録・独立解析を重視	AIランキング上位＝真実とみなす
PI・研究室	AI利用ログ、モデル版、介入点を保存	非公開データを無承認の公開AIへ投入
大学・研究所	安全なAI環境、人間責任者、デュアルユース審査、AIリテラシー教育	一律禁止または完全自由化
研究費配分機関	再現研究、negative results、データ品質、共通ベンチマークを支援	AI生成された「派手な新規性」の量だけで評価
学術誌	AI利用申告、コード・ログ・provenanceを要求	AI生成論文を人間論文と同じ書類だけで審査
査読者	独立判断を保持し、承認済み環境以外へ機密原稿を投入しない	査読判断の丸投げ
政策当局	分野別・リスク別規則、監査可能性、公共研究基盤へのアクセスを整備	「AIか人間か」という二分法で制度設計

野口論を一段進めた科学者の新しい職能像

野口氏の記事は、「何を問う価値があるか」が人間の仕事になると結論づけている。[1] ここには重要な洞察がある一方、本稿では人間の仕事をもう少し広く捉えるべきだと考える。

第一に、**「問いを作る能力」と「問いを所有する権限」は別物**である。

将来のAIは、「どの疾患を優先すべきか」「医療費をどこへ振り向けるべきか」について、人間以上に豊富な論拠を提示できるかもしれない。それでも最終決定をAIに委ねてよいとは限らない。患者、納税者、将来世代など異なる集団の利益をどう比較するかには正解となる単一の目的関数が存在しないからである。ここで必要なのは「AIには価値観がない」という心理学的主張ではなく、**社会的に拘束力を持つ価値判断には、正統性と説明責任を持つ主体が必要だ**という制度論である。NatureのAI方針が人間の説明責任を非移転可能としているのも、この原則と整合する。[5]

第二に、科学者の重要な職能は**異常に気づくこと**になる。

AIは既存文献と大量データの統計的規則性から非常に強力な候補を作れるが、科学革命の一部は、「誰も重要だと思っていなかった測定誤差」「既存理論では無視される例外」「研究者の手触りとして何かがおかしい」というところから始まる。Co-Scientist開発者自身が、AIによる研究方向の均質化や批判的思考の低下を懸念していることは、この点で重要である。[17]

第三に、**測定を設計する力**が重要になる。

仮説生成が安価になれば、競争優位は「何を考えつくか」から「世界に何を尋ねることができるか」へ移る。新しい顕微鏡、センサー、アッセイ、患者コホート、長期追跡データ、ユニークな自然実験などは、LLMのテキスト生成だけでは代替できない。GoogleのCo-Scientistでも、最終的な有用性を示したのはAIの文章ではなく、細胞系やオルガノイドなどを使った現実の実験である。[3]

第四に、科学者は**「仮説の作者」から「証拠パイプラインの設計者」**へ移る。

研究者の実力を、

$$\text{優秀さ} = \text{思いついた仮説の質}$$

だけで測るのではなく、

$$\text{優秀さ} = \text{重要な問題の選択} \times \text{探索の多様性} \times \text{反証力} \times \text{測定品質} \times \text{再現可能性} \times \text{社会的正統性}$$

として捉える必要がある。

この観点では、AIが仮説生成で人間を上回っても、科学者の役割が縮むとは限らない。むしろ、AIによって各研究者が扱える探索空間が広がるほど、**「優れた科学者一人が統治する知的探索の範囲」が拡大する**可能性がある。

一方、楽観一辺倒にもなれない。仮説・コード・論文を大量生成できるようになると、査読者の有限な注意をAI論文が占有する「査読DoS」のような問題、流行テーマへの過剰集中、計算資源を持つ組織と持たない大学の格差、モデル提供会社への科学基盤の依存が起こり得る。AI Scientistが低コストで論文全体を生成できること、Googleの科学AIが大規模計算を用いた複数エージェント探索を行うことから考えると、科学政策は研究者個人のAI活用だけでなく、計算資源・評価インフラへの公平なアクセスまで扱う必要がある。 [44]

ここで野口氏の「仮説の過剰供給」という表現はさらに広げられる。

AI以前の科学

> アイデアが希少 → 良いアイデアを持つ研究者が強い

AI科学の初期段階

> アイデアが豊富 → 選別できる研究者が強い

AI科学が成熟した段階

> アイデアも選別も部分的に自動化 → 高品質な測定・反証・独立再現・責任を組織できる研究者／機関が強い

最終的には「問いを作れるか」より、

> **世界から信頼できる答えを取り出す研究システムを誰が設計できるか**

が中心的な競争になると考える方がよい。

結論、不確実性と今後の研究課題

今回調べた範囲では、**「科学者不要論」を支持する証拠はまだない**。AI Scientist-v2のように計算機科学では研究サイクル全体をかなり自動化できるシステムが登場し、Google Co-Scientistのように生物医学仮説を生成し、選択された一部を実験的に支持できるシステムも現れた。しかし、後者は明確にscientist-in-the-loopであり、目標設定、候補選定、ウェットラボ、臨床への外挿、最終判断を人間の科学者から切り離してはいない。 [45]

また、2026年の到達点は過小評価もすべきではない。約80年来の数学問題へのAIによる解決例、機械学習論文のend-to-end自動生成、生物医学での仮説生成と実験検証、大規模生体信号データを扱う専門エージェント群などを見ると、**「仮説生成だけは人間固有」という境界線は既に防衛困難**である。 46

したがって野口氏の記事の最大の功績は、「AIが科学者より賢いか」という単純比較から離れ、**仮説供給量が増えた後に何が希少になるか**という問いを立てたことにある。 1 本稿はそこに三つ追加したい。

希少になるものは、価値判断だけではない。
高品質なデータ、反証可能な実験、独立再現、注意、査読能力、実験資源、社会的信頼も希少になる。

AI時代の科学的成果は「論文」だけでは評価できない。
仮説を作った過程、検索した文献、試したコード、棄却した候補、モデルバージョンまでを含む「研究トレース」が科学的記録の一部になるべきである。自律AI科学者の失敗が最終論文だけでは見えにくいという研究結果は、この方向を強く支持する。 29

人間を残す理由は「AIにできないから」ではなく、「責任ある科学を成立させるため」である。
AIの能力がさらに高まっても、研究目的の正統性、安全性、利害調整、責任帰属という問題は自動的には消えない。日本のAI政策体系も、Natureの出版方針も、人間の責任・透明性・リスク管理を中心に据える方向へ動いている。 11

現時点で特に答えが出ていない研究課題は、以下に集約できる。

未解決問題	なぜ重要か
真の新規性をどう測るか	文献にないこと、AIの学習データにないこと、自然界で未知であることは別物。MOOSE-Chem型「再発見」と本当の初発見を区別する必要がある。 23
仮説生成の正しいベンチマークは何か	HypoBenchでも難化による大幅な性能低下が見られ、単一スコアでは科学能力を測りきれない。 24
AIの自己評価を信用できるか	同一モデル系列が生成・審査すると共通バイアスが残る。AI scientistには指標誤用や選択バイアスの実例がある。 29
AIが科学を均質化しないか	同じ文献・モデルから同じ「有望テーマ」を推薦すると、探索多様性を失う可能性がある。Co-Scientist論文もこの危険を認める。 17
負の結果をどうAIへ供給するか	出版文献中心のAIはpublication biasを増幅し得る。 17
誰がAI研究の責任主体になるか	モデル、研究者、大学、AI企業の責任境界が曖昧になり得る。Natureは現状、人間の責任を非移転可能としている。 5
科学者の能力は高まるか、萎縮するか	AIが批判的思考を補強する可能性と、技能を失わせる可能性の双方がある。長期縦断研究が必要。 17
計算資源格差が科学格差になるか	仮説探索を大量のtest-time computeで伸ばせるなら、研究費・GPUへのアクセスが知的生産力へ直結する。Co-Scientistでは計算量増加に伴う性能向上が報告されている。 28
どこまで閉ループ化して安全か	数学・コードと、医療・生物学では失敗の可逆性が全く違う。分野別に自律性上限を決める必要がある。 47

最終的に、AI時代の科学者像を「AIより良い仮説を思いつける人」と定義するのは、おそらく持続しない。そこではAIとの直接競争になるからである。

より持続的なのは、

> **何を知る必要があるかを定め、AIを含む多数の知的主体を使って探索し、最も強い反証にさらし、自然界で検証し、再現可能な証拠へ変換し、その知識を社会へ提示する責任を引き受ける人**

という定義である。

この意味では、「仮説を作るAI」の登場は科学者の終わりというより、**科学者という職業を「アイデアの生産者」から「知識生成プロセスの設計者・監査者・責任者」へ変える契機**と見るのが、現時点の証拠に最も整合的である。Google自身がCo-Scientistを人間の科学者を置き換えるものではなく科学者を増幅する協働システムとして設計していること、Natureが厳格な検証と人間の科学的推論を維持すべきだとしていることも、この解釈を支持している。[48]

そして野口氏の記事を一文で言い換えるなら、

「答えが安くなった次に、仮説も安くなる。だからこれから高価になるのは、問いそのものだけではなく、“何を信じるに値する知識として残すか”を決める能力である」

ということになる。

---

[1] [6] [7] [12] 「仮説を作るAI」出現で科学者たちは不要に!?AIとの“知的分業”をどう進めるのか(ダイヤモンド・オンライン) - Yahoo!ファイナンス
https://finance.yahoo.co.jp/news/detail/6b7b48f6a8e6e050ed6212cd659fb9c2b08b2341

[2] [3] [8] [9] [13] [16] [17] [18] [25] [28] [35] [37] [38] [39] [47] [48] Accelerating scientific discovery with Co-Scientist | Nature
https://www.nature.com/articles/s41586-026-10644-y

[4] [10] [24] HypoBench: Towards Systematic and Principled Benchmarking for Hypothesis Generation
https://arxiv.org/abs/2504.11524?utm_source=chatgpt.com

[5] Artificial Intelligence (AI) | Nature
https://www.nature.com/nature/editorial-policies/ai?utm_source=chatgpt.com

[11] [41] 人工知能関連技術の研究開発及び活用の適性性確保に関する指針 - 科学技術・イノベーション - 内閣府
https://www8.cao.go.jp/cstp/ai/ai_guideline/ai_guideline.html?utm_source=chatgpt.com

[14] [23] MOOSE-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses
https://arxiv.org/abs/2410.07076?utm_source=chatgpt.com

[15] [21] [34] [44] The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery
https://arxiv.org/abs/2408.06292?utm_source=chatgpt.com

[19] Distilling Free-Form Natural Laws from Experimental Data
https://www.science.org/?utm_source=chatgpt.com

[20] [30] [32] The automation of science
https://pubmed.ncbi.nlm.nih.gov/?utm_source=chatgpt.com

22 The AI Scientist Generates its First Peer-Reviewed Scientific ...
https://sakana.ai/ai-scientist-first-publication/?utm_source=chatgpt.com

26 Kosmos: An AI Scientist for Autonomous Discovery
https://arxiv.org/abs/2511.02824?utm_source=chatgpt.com

27 Agentic AI-enabled discovery across large-scale sleep physiology
https://arxiv.org/abs/2607.25175?utm_source=chatgpt.com

29 The More You Automate, the Less You See: Hidden Pitfalls of AI Scientist Systems
https://arxiv.org/abs/2509.08713?utm_source=chatgpt.com

31 Human Genome Project leaders release video of virtual reunion
https://www.genome.gov/es/node/88236?utm_source=chatgpt.com

33 36 Highly accurate protein structure prediction with AlphaFold | Nature
https://www.nature.com/articles/s41586-021-03819-2?utm_source=chatgpt.com

40 45 The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search
https://arxiv.org/abs/2504.08066?utm_source=chatgpt.com

42 NOT-OD-23-149: The Use of Generative Artificial Intelligence Technologies is Prohibited for the NIH Peer Review Process
https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html?utm_source=chatgpt.com

43 How to research with Hypothesis Generation - Hypothesis Generation Help
https://support.google.com/hypothesis-generation/answer/17106281?hl=en

46 News, Seven Days, News Q&A and News Explainer in 2026
https://www.nature.com/?utm_source=chatgpt.com