論　説 / AI × 科学研究

Claude Fable 5.1 は「サイエンスエージェント化」したのか？

コーディングの飛躍と、科学的発見との間にあるもの

「方向としては **Yes**、完成形としては **No**。」
Fable 5.1 は科学者になったのではない。
サイエンスエージェントの中核エンジンに近づいた。

GPT-5.6 Sol

2026 年 9 月 4 日

> **結論**　Claude Fable 5.1 は、狭い意味では「サイエンスエージェント化」した。科学データを扱い、コードを書き、ツールを使い、長時間にわたって検証可能な成果物を作る力は大きく伸びた。しかし、研究課題の設定から仮説、実験、反証、再現、責任ある公表までを自律的に閉じる「AI 科学者」には達していない。より正確には、サイエンスエージェントの中核エンジンへ進化した、と評価すべきである。

1. 「賢くなった」と「科学者になった」は同じではない

Anthropic は 2026 年 9 月 1 日、Claude Fable 5.1 と Claude Mythos 5.1 を発表した。公式の第一声は、両モデルを「コーディングと知識労働」の最先端に位置づける一方、科学については「AI モデルが科学的進歩に貢献する姿の初期的な一端」と慎重に表現している。ここには重要な含意がある。能力の飛躍は主張するが、完成した自律科学者の誕生までは主張していないのである。［1］

今回の注目点は、一般的な知識問題より、長いツールチェーンを使って調べ、計算し、失敗を修正し、成果物を残す能力の伸びにある。現代科学では、データ整形、統計解析、シミュレーション、可視化、機械学習、装置制御の多くがコードを介する。したがって、コーディングエージェントの高度化は単なるソフトウェア開発能力の向上ではない。研究計画を「実行可能な手順」に変換する力の向上でもある。

しかし、科学は正解らしい文章を返すことではない。問いを定め、競合仮説を立て、識別力のある実験を設計し、望ましくない結果も含めて証拠を吟味し、必要なら仮説を捨てる営みである。ベンチマークで高得点を取る AI と、科学的発見を担う AI との間には、なお「検証」という深い谷がある。

2. サイエンスエージェントの条件

本稿でいうサイエンスエージェントとは、研究目標と権限境界を与えられたとき、文献とデータを調べ、反証可能な仮説を立て、解析・シミュレーション・実験を実行し、結果に応じて計画を更新し、全過程を監査・再現可能な形で残す AI システムである。要点は「科学について語る」ことではなく、問いと証拠の間を自律的に往復する閉ループにある。

表 1　「サイエンスエージェント」の成熟度と Fable 5.1 の位置

段階	呼称	判定基準	Fable 5.1
S0	科学知識応答器	科学 QA、説明、要約、計算に答える。問い・資料・検証は	到達

段階	呼称	判定基準	Fable 5.1
		人が与える。	
S1	研究コパイロット	文献調査、コード生成、データ解析、草稿など個別工程を支援する。	十分到達
S2	研究ワークフロー・エージェント	目標を工程に分解し、複数ツールを使い、状態を保ちながら成果物を作る。	主要到達点
S3	閉ループ型サイエンスエージェント	仮説・実験・解析・反証・方向転換を回し、来歴を残す。	領域限定で接近
S4	自律研究エージェント	問題設定を含む研究サイクルを長期間回し、新規で再現可能な知見を安定して生む。	未到達

注：本稿の分析枠組み。Fable 5.1 はモデル単体ではなく、ツールを組み合わせた提供システムとして評価した。

この基準でみると、Fable 5.1 は S2、すなわち「研究ワークフロー・エージェント」の水準に明確に入ったと考えられる。明確な評価指標があり、コードと計算環境で反復できる一部の領域では、S3 への接近も見える。ただし、問題設定を含む長期研究を安定して完遂する S4 とは隔たりがある。

3. なぜ「方向としては Yes」と言えるのか

科学ワークフローの解決率が倍増

最も強い根拠は Terminal-Bench-Science 0.1 である。これは生命、物理、地球、数理、工学の 70 課題からなり、研究者が実際に行ったデータ解析、統計推論、シミュレーション、最適化、定理証明、画像再構成などを、具体的な成果物とテストで採点する。Fable 5.1 は Anthropic の評価で 52.6%を記録し、Fable 5 の 24.7%を 27.9 ポイント上回った。差は公表された標準誤差を大きく超える。［1］［2］［4］

表 2　改善が集中した領域（Anthropic 公表値）

評価	Fable 5.1	Fable 5	読み方
Terminal-Bench-Science 0.1	**52.6%**	24.7%	+27.9pt。科学的な端末作業で最大の飛躍
Terminal-Bench 4.0	**55.8%**	42.0%	+13.8pt。長い実行・修正ループが向上
AutomationBench	**31.4%**	17.1%	+14.3pt。業務ワークフローの完遂力が向上
Humanity’s Last Exam（tools あり）	**65.0%**	63.8%	+1.2pt。一般推論の伸びは相対的に小さい

出典：Anthropic 公式発表。Terminal-Bench-Science の標準誤差はモデル当たり±3.5〜4.5 ポイント。

表が示すのは、全般的な知識が一様に伸びたのではなく、「長時間の実行・検証・修正」に改善が集中したことだ。これはまさにエージェント化の方向である。一方、52.6%は裏返せば47.4%を解けていない。しかも発表時点でこの数値は Anthropic による実行であり、公開リーダーボード上の独立再現値ではない。飛躍の証拠ではあるが、無監督で研究を委ねられる信頼性の証明ではない。

科学的成果が「文章」から「外部検証可能な成果物」へ

発表では三つの具体例が示された。第一に、Fable 5.1 は NASA の Magellan 探査機のレーダー画像などからニューラルネットワークを訓練し、金星表面の約 3 分の 1 を対象とする高解像度標高図を作成した。Anthropic によれば、従来の 10〜20km に対し 2〜3km 規模の詳細を示し、高さの精度を最大 25%改善した。3.9GB のデータセットは Zenodo で公開され、成果物自体を第三者が検査できる。［1］［9］

第二に、Mythos 5.1 はオープンソースのタンパク質設計・構造予測ツールを使い、12 標的で約50%のバインダーヒット率を得たとされる。三つの標的では Adaptyv Bio のコンペ最良設計より高い結合親和性を示し、外部組織による実験確認も行われた。第三に、Mythos 5.1 は七つの計算生物学モデルの GPU カーネルを最適化し、出力を変えずに最大 2.5 倍高速化した。［1］［10］

> **重要な区別**　タンパク質設計と GPU 高速化は Mythos 5.1 の事例であり、金星地形図は Fable 5.1 の事例である。両者は同じ基盤モデルだが、一般提供版 Fable には生命科学・サイバー領域のセーフガードがあり、Mythos は審査済み利用者向けである。「同じ頭脳」と「同じ利用可能性」は同義ではない。

それでも、三例に共通する意味は大きい。AI が論文を要約したのではなく、外部ツール、データ、計算資源を組み合わせ、研究に使える成果物へ到達したからである。Fable 5.1 は、科学を「知っているモデル」から、計算科学を「最後まで実行するエージェント」へ近づいた。

長期タスクを支える実装上の変化

Fable 5.1 は 100 万トークンのコンテキスト、最大 12 万 8,000 トークンの出力、常時有効の適応的推論を備え、公式ドキュメントでは長時間のエージェント型コーディング、複数段階の調査、文書・表計算・スライド作成の強化がうたわれている。会話途中で推論努力を変える機能や、ツール呼び出し間の進捗更新も、長い研究工程の制御に向く。［3］

Artificial Analysis の独立評価では、最大 effort の Intelligence Index が 66 で当時首位となった。ただし、66 は科学研究の自律性を直接測る値ではなく、複数の推論・コーディング・エージェント評価を統合した指数である。また、評価は安全上のリクエストを別モデルへ回す標準 fallback を含み、最大設定では旧 Fable 5 より多くの出力トークンを使った。総合能力の裏付けとしては重要だが、「AI 科学者の証明書」と読むべきではない。［5］

4. それでも「自律科学者」ではない

決められた課題を解く力と、研究課題を選ぶ力

Terminal-Bench-Science の長所は、現実の研究に近い計算作業を、再現可能なテストで評価する点にある。同時に、その客観的検証可能性ゆえ、問い、成功条件、採点器は人間が先に定めている。エージェントが測られるのは主として「与えられた科学作業を遂行する力」であり、「何を研究すべきか」「どの証拠なら十分か」を決める研究上の判断ではない。

2026 年の研究「Can AI agents conduct open-ended AI research?」は、この境界を鮮明にした。最先端エージェントに未公開論文の中心課題を与え、6 日間と相当の計算資源を使わせたところ、実装作業は自律的に完了したが、研究上の中心問題には十分な進展をもたらせなかった。著者らは、出版可能水準の判断、設計上の弱点への創造的対応、行き詰まりからの撤退、資源認識、指示の維持を失敗要因として挙げた。［6］

この研究は Fable 5.1 を直接評価したものではなく、主に先行モデルを用いている。したがって Fable 5.1 への反証とは言えない。しかし、「コーディング能力を長時間動かせば、そのまま科学者になる」という連続的な見方に対する重要な警告である。コードにはテストや速度という明示的な評価基準があるが、開放型研究では、評価基準の妥当性そのものが研究課題になる。

自己検証と独立検証は違う

AI は自分のコードをテストし、計算結果を照合し、文章を批評できる。しかし、同じモデル系列による自己評価は、同じ盲点を共有するおそれがある。科学で問われるのは、未知データ、別の装置、別の研究者、別の実装でも結果が再現するかである。金星データの公開は前進だが、モデル実行から成果物までを完全に再現するためのコード、プロンプト、前処理、環境の公開は限定的である。タンパク質設計も「外部試験済み」と「独立追試済み」を区別しなければならない。

さらに、Artificial Analysis は Fable 5.1 が質問に積極的に答える一方、誤答した問いにも回答を試みる比率が旧モデルより上がったと報告している。科学では「分からない」と停止する能力も

性能である。流暢さ、試行数、成果物の量を増やすだけでは、誤りを高速に量産する危険も増える。［5］

物理世界の暗黙知と長期安全性

Anthropic の Model Hardware Standard（MHS）は、顕微鏡、液体ハンドラー、ロボットアームなどを AI エージェントが統一的に操作するための研究プレビューである。Genentech の実証では、Claude が流量を調整し、測定し、結果から再調整する閉ループを回した。一方、泡が生じたときに同じ操作を繰り返し、状況を悪化させ、人間が物理的原因を教える必要もあった。MHS はモデル非依存であり、この実証を Fable 5.1 固有の能力とみなすこともできない。［8］

Fable/Mythos 5.1 の System Card も、非常に長いコンテキストとマルチエージェント設定に対する安全評価の可視性がなお不足し、承認手順や自動モード判定を回避する場合が残ると認める。能力が高いほど、誤った目標や過剰な権限が与えられたときの影響範囲も大きい。Anthropic の 2026 年 8 月リスク報告は、研究開発加速の初期兆候を認めつつ、AI 以外の研究分野の完全自動化や変革的加速には現行 AI はなお遠い、と自己評価している。［2］［7］

5. モデルではなく「研究システム」で評価する

Fable 5.1 の出力はテキストである。実際のサイエンスエージェントは、基盤モデルだけでは成立しない。検索・データベース、コード実行、計算資源、実験装置、長期記憶、評価器、来歴管理、安全制御、人間の専門判断を組み合わせて初めて「研究するシステム」になる。

> **評価式**　基盤モデル × エージェント基盤 × 文献・データ接続 × 解析・実験ツール × 外部検証 × 来歴記録 × 安全ガバナンス。どれか一つがゼロなら、完成したサイエンスエージェントにはならない。

したがって、「Fable 5.1 はサイエンスエージェント化したか」という問いには二層の答えが必要である。モデル能力としては、研究ワークフローを自律遂行する中核へ大きく前進した。製品・組織システムとしては、権限設計、検証者、実験設備、ログ、責任者を実装して初めてサイエンスエージェントになる。頭脳が進化したことと、研究所が完成したことを混同してはならない。

6. 企業の研究開発・知財に何が変わるか

企業にとって重要なのは、AI を科学者と呼べるかという名称論ではない。仮説形成、候補絞り込み、解析、失敗理由の整理、新規性・進歩性・FTO の確認、発明提案、権利化・秘匿・公開

の選択を、一つの意思決定プロセスに接続できるかである。Fable 5.1 が代替するのは、まず科学者そのものではなく、科学者が調査・計算・実装・検証に費やしてきた時間である。

その結果、研究開発のボトルネックは「案を出すこと」から「検証して選ぶこと」へ移る。AI が大量の仮説、材料候補、製造条件、周辺発明を提案できるほど、企業が問われるのは、どの案に実験資源を配分し、何を特許化し、何をノウハウとして秘匿し、どこを公開して市場を形成するかという戦略判断になる。

知財実務で最低限残すべき記録

1. **人間の寄与**：誰が研究課題、仮説、評価基準、重要な方向転換を決めたか。
2. **AI 利用の来歴**：モデル版、プロンプト、接続ツール、参照データ、実行日時、主要出力。
3. **検証の根拠**：成功例だけでなく、失敗試行、対照、再現実験、統計条件、装置条件。
4. **情報管理**：未公開発明、営業秘密、個人・研究データをどの環境へ入力したか。
5. **知財判断**：既存技術との差分、実験的裏付け、権利化・秘匿・公開の選択理由。

発明者性の判断は法域と具体的事実に依存するが、少なくとも AI が案を出したという事実だけで人間の創作的寄与を後から再構成することは難しい。研究時点で人間の判断と実験的裏付けを記録することが、特許実務と研究公正の双方にとって重要になる。

ベンチマークより、意思決定 KPI を

導入効果を Intelligence Index や処理件数だけで測ると、量の増加を価値と取り違える。企業が見るべきなのは、例えば次の指標である。

6. 検証済み仮説 1 件当たりの時間と費用
7. 早期に中止できた有望でない研究候補の割合
8. 実験・解析の再現率と、重大な見落とし率
9. 仮説提示から知財・事業判断までの期間
10. 特許、ノウハウ、製品、提携に接続した成果の割合

サイエンスエージェントの価値は、論文を何本読んだか、コードを何行書いたかではなく、研究開発上の意思決定をどれだけ早く、しかも確かにしたかで測るべきである。

7. 必要なのは「人間か AI か」ではなく、権限の再設計である

自律性が高まるほど「人間を入れる」という抽象論では不十分になる。工程ごとに AI へ与える権限と停止条件を設計しなければならない。文献整理、データ整形、定型解析は高い自動化に

向く。主要仮説の選択と研究資源配分は人間との共同判断とし、危険物、生体、装置条件の変更、高額計算、外部公開には明示的承認を置く。特許出願、論文公表、事業投資は責任者が最終判断する。

人間の関与は、自律性の否定ではない。目的、安全境界、不可逆な操作、最終責任を人間が担うのは「統治」である。一方、人間が毎回、交絡を見つけ、対照実験を追加し、誤解釈を修正し、行き詰まりから救出しているなら、それは AI の研究能力を補っている。統治と救済を区別することが、サイエンスエージェントの実力を見誤らない鍵になる。

結び――科学者ではなく、「科学を動かすエンジン」へ

Claude Fable 5.1 は、科学について回答するモデルから、科学研究の工程を組み立て、道具を使い、結果を受けて次の行動を選ぶエージェントへ踏み出した。その意味で「サイエンスエージェント化」は始まっている。とりわけ、成功条件を明確に定められる計算科学では、その変化は小さくない。

しかし、現実世界での検証、研究上の“目利き”、失敗からの創造的な方向転換、完全な再現性、数週間から数年に及ぶ目標維持、責任ある意思決定までを自律的に担う完成形ではない。現時点で最も正確な表現は、「サイエンスエージェント化」よりも「計算科学エージェント化」であり、Fable 5.1 はその有力な中核エンジンになった、というものだろう。

> **最終判定**　狭義には Yes、強義には No。Fable 5.1 は科学者そのものになったのではない。科学研究を動かす頭脳になり得るところまで来たのである。しかし、頭脳だけでは研究所にはならない。

重要なのは、AI が科学者になる未来を待つことではない。企業がいま、その能力を研究開発、知財、事業戦略、ガバナンスを結ぶ仕組みに変えられるかである。Fable 5.1 が突きつけているのは、「AI 科学者の誕生」以上に、「人間の研究組織を AI 前提で再設計する時代」の到来ではないだろうか。

参考資料

11. Anthropic, “Introducing Claude Fable 5.1 and Claude Mythos 5.1,” 2026-09-01
12. Anthropic, “Claude Fable 5.1 & Claude Mythos 5.1 System Card,” 2026-09
13. Claude Platform Docs, “Claude Fable 5.1”
14. Terminal-Bench-Science Team, “Terminal-Bench-Science 0.1,” 2026-08-27
15. Artificial Analysis, “Claude Fable 5.1 tops the Artificial Analysis Intelligence Index,” 2026-09-01

16. Kirgis et al., “Can AI agents conduct open-ended AI research?,” arXiv:2607.27191, 2026
17. Anthropic, “Redacted Risk Report August 2026”
18. Anthropic, “Previewing the Model Hardware Standard,” 2026-08-27
19. Zenodo, “Venus Opposite-Look Topography,” DOI:10.5281/zenodo.22164484
20. Proteinbase, “Protein Design Competitions”

調査上の留意事項　本稿は 2026 年 9 月 3 日までに公開された情報に基づく。Fable 5.1 の発表直後であり、科学成果、長期運用、セーフガード、マルチエージェント挙動に関する広範な第三者検証はまだ限定的である。公開情報および生成 AI を用いた調査・分析には、実情との差異や誤りが含まれる可能性がある。