科学 AI の自律性レベル

Levels of Autonomy in Science Automation

高橋恒一 [123*]
Koichi Takahashi

[1] 理化学研究所　科学研究基盤モデル開発プログラム
[2] 慶應義塾大学大学院　政策・メディア研究科
[3] AI アライメント・ネットワーク

Abstract: Artificial Intelligence (AI) in scientific and technological research is anticipated to be one of the fields with the most significant ripple effects among long-term AI applications. When AI is viewed as a form of automation technology, its utility is proportional to the extent to which it can reduce or simplify human operation and instruction. Consequently, the efficacy of scientific AI is considered to be closely related to its autonomy. This paper proposes a seven-level autonomy scale, ranging from 0 to 6, for scientific AI, primarily focusing on experimental science fields. This proposal draws inspiration from the autonomy levels for automated driving systems as defined by the Society of Automotive Engineers (SAE). This article discusses the definition of each level, the milestones for their realization, and their potential impacts on academia, technology, and society.
keywoards: artificial intelligence, automation of science, AI-driven science

要旨： 科学技術研究は、長期的な AI 応用の中でも最も波及効果が大きいと見込まれる分野の一つである。AI を自動化技術と捉えた場合、その効用は人間の操作や指示の削減あるいは簡略化の程度に比例する。したがって、科学 AI の効用はその自律性と密接に関連すると考えられる。本稿では、米国自動車技術会（SAE）が提示した自動運転のレベル分類にヒントを得て、実験科学分野における科学 AI の自律性を 0 から 6 までの 7 段階で設定することを提案する。各レベルについて、その定義、実現に向けたマイルストーン、および学術、技術、社会へのインパクトについて論じる。
キーワード： 人工知能、科学の自動化、AI 駆動科学

0 序文

本稿は、著者が 2020 年 2 月 26 日から 27 日にかけて Alan Turing 研究所で開催された AI Scientist Grand Challenge ワークショップ[１]に参加した際に構想し、その後理化学研究所未来戦略室において「科学 AI の自律性レベル設定」として 2021 年から 2022 年にかけて書き上げたレポートに基づいたものである。なお、2022 年以降の分野の進展、およびいくつかの発表機会[２][３][４]や私信で得られたフィードバックを踏まえた改訂を行っているが、科学 AI の自律性レベルの設定自体は変わっていない。

1 はじめに

人工知能の科学研究への応用はチューリング以来古くから目指されてきた。過去の取り組みには、定理証明のように有限の公理系から明らかに論理推論が可能な系への適用、投薬計画などの特定のドメインにおける知識工学的アプローチ、細胞内反応経路上での定性推論の応用などがあったが、いずれも出発点となる知識とその表現方法を人間が与える点に限界があった。近年の計算力の爆発的な発展を背景とした深層生成モデルや予測符号化モデルなどによる教師なし学習の発展により、データ駆動で適切な潜在変数の構成や変数間の関係を自動で抽出することが取り組まれている。また、情報学の整備により領域知識の表現方法の整備と蓄積が進み個別事例を超えた知識の構造化が可能になりつつある。さらに、近年の自動実験ロボット技術の急速な普及により、高品質のデータを大量に生成し学習を駆動する見込みも立ちつつある[５][６][７]。

本稿では、まず、現在から将来にわたって科学・技術研究の自動化の度合いがどのように進展してゆくのかを考えるうえでの参照点の一つとして、「科学 AI の自律性レベル」を設定し、その定義を述べる。次に、主に実験科学分野を念頭に、自律性レベルの各段階における技術的要件や実現条件などをマイルストーンとして整理する。さらに、自律性レベル設定の各段階における技術的、学術的、および社会的インパクトを議論する。

なお、本稿は具体的な実現時期や年代を示した未来予測を目的としたものではない。また、本稿における考察は、AI による科学の自動化が逐一の人間の介入を必要とする低い自律性のレベルから最終的には人間の介入を必要としない高い自律性のレベルへと発展してゆくという仮説に基づいている。

2 定義

ここでは、科学 AI の自律性レベル設定を提案する。科学 AI を一種の自動化装置と考えた場合、目的が自動化であるがゆえに、その価値は人間が装置に与えなければいけない指示の頻度と直接性が下がると下がるほど高まると考えられる。これは、例えば自動車の自動運転を考えた場合に、人間が直接ブレーキとアクセルを操作するよりもオートクルーズシステムで速度を一定に保つ仕組みがあったほうがより有用であること、また、行き先を入力するだけで運転の全操作を機械に任す選択肢があったほうが運転者にとってはさらに価値が高いということと同じである。

自動車の自動運転では、米国の SAE（自動車技術会）がレベル０から５までの６段階で基準を示している[４]。レベル０は運転自動化なし、１は運転支援、２では部分運転自動化で、いずれも運転主体は人となっている。レベル３以降では運転主体がシステムに移り、レベル３では高速道路などの特定の運転領域においてシステムが全ての運転操作を行うが人による適切な介入を必要とする条件付き運転自動化、レベル４では特定条件下でシステムが全ての操作を実施する高度運転自動化と定義されている。レベル５は運転領域を限定せず、常にシステムが全ての運転操作を行う完全自動運転であり、人は行き先を指定するだけで自動車が運行される。つまり、全体として、自動化技術が全く導入されていないレベル０から運転動作から完全に人の介入を排除したレベル５までの両極端の間に４つの中間的レベルが設定された構造となっている。

図 1 に今回提案する科学 AI の自律性レベルを示す。SAE の自動運転レベルと同様に、最も自律性が低いレベル０では計算機が完全に人間の道具として用いられ、最も自律性が高いレベル６では一切の人間の介入が必要とされず研究目的の設定から遂行までの全てのプロセスが自動化される。これらの両極端であるレベル０とレベル６のあいだにいくつの中間的レベルをどのような順番で設定するのが適切なのかは自明ではなく、恣意性が生じる。そのため、

Lv.	機械が行うこと	主要な入力	説明
0	計算	・プログラム	データ処理、分類、シミュレーションなど
1	閉鎖系の解探索	・探索空間 ・解の評価関数	定理証明、囲碁、データ解釈、医療診断など
2	開放系の探索	・探索範囲 ・解の評価関数	物質探索、細胞培養条件探索など 実験計画、ブラックボックス最適化、法則発見、非モデルベースアクティブラーニング等
3	形式固定の仮説生成	・モデルの表現方法 ・初期モデルの生成方法 ・獲得情報の評価関数	機械は検証が所与の獲得情報価値関数を最大化するような新たな知識候補となる仮説命題を、人間が与えられた形式内で生成し、自ら実験計画を立ててそれを検証する。
4	形式自由の仮説生成	・研究対象 ・獲得情報の評価関数	レベル３と同様だが、生成される仮説の形式自体も探索対象とする。検証に新たな観測量や観測装置が必要となる仮説および実験計画の生成がありうる。
5	研究立案	・研究目的 （検証すべき仮説）	与えられた研究目的に従った獲得情報価値関数自体の提案、改善。
6	研究目的の設定	なし	研究の対象とその目的自体を設定し、研究を遂行。

図 1　科学 AI の自律性レベル

何のためにレベル設定を行うのかという目的の設定が重要である。このようなレベル設定の目的としては、今後の技術発展の予測に資することや、社会的インパクトの予測と整理などがある。今回の提案では、今後の技術的な発展の道筋を示すロードマップとしての用途を中心に考慮しつつ、それぞれの社会的インパクトについても議論しやすい中間レベル設定を目指した。

定義：レベル０

レベル０は計算機が計算機として用いられ、自律性を発揮していないケースである。例えば、世界最初の大型計算機である ENIAC が 1946 年に砲弾の弾道計算をした時行ったのは人がスイッチと配線盤で示したプログラムを忠実に実行しただけであり、計算機は何ら自律性を発揮していない。データ処理やシミュレーションなどの計算も、その内容がいかに高度で大規模であっても、指示された司令を実行しているということには変わりなく、自律的ではない。深層学習モデルによる画像の分類なども、分類というタスクの実行にのみ注目して考えれば、自律的ではないだろう。

定義：レベル６

レベル１から５の中間的なレベルを説明する前に、まず完全に自律的な極端であるレベル６を説明する。レベル６では、AI 自身が研究目的を設定しそれに適した研究対象を見つける、ないしは逆に、ある対象に興味を抱き、それに対して研究目的を設定し、研究を遂行する。自律的に研究目的を設定するということは、無際限な可能性の中から特定の研究対象を選別するということであり、この点で AI は独立して価値判断を行い行動する自発性を持っている。好奇心を持つことも価値判断の一形態であると考えれば、特定の対象への興味を出発点にして、それに従属的に研究目的を設定する場合にも同様のことが言える。

定義：レベル１

自律性レベル１は中間レベルの最初の段階である。自律性レベル１の科学 AI は、閉鎖系で解探索を行う。ここで言う閉鎖系とは、探索を適用する対象の内部状態がそれ自身の内的な要因により完全に決定され外的な要因により変化しないことを意味し、もし対象の内部状態が計算機内で完全に表現可能な場合にはこの特性は多くの場合純粋に計算機内での計算による探索を可能にする。例えば、推論システムの一種である自動定理証明システムの多くは、公理と推論規則のみから探索ヒューリスティクスを用いて命題を証明するため、探索がコンピュータ内で完結しており、計算の途中である解の評価値が外的な要因により変化しないという意味で閉鎖系である。

囲碁や将棋などは完全情報ゲームであり、ゲームのルールや過去の全プレイヤーの打ち手の情報は全て与えられた状況で次の打ち手を探索するため、次の最適な打ち手を探す探索がコンピュータ内で完結している。このようなシステムは、対象とする問題の難しさにもよるが、既に多くが実用化されている。例えば囲碁や将棋などのゲームで人間のチャンピオンを打ち負かす性能を持つ AI は既に存在する。また、スタンフォード大学が６０年代に開発した Dendral は有機化学の知識を用いて質量分析の結果から未知の有機化合物を同定するエキスパートシステムである[９]。同大学は Dendral を発展させ、７０年代に Mycin と呼ばれる伝染性の血液疾患に対する抗生物質投薬計画を立案する AI システムも開発した[１０]。Dendral と Mycin は、いずれも人間の入力に対して解候補を提示する構造になっており、もし解候補を仮説の一種であると考えれば後述の自律性レベル３の議論で取り上げるような仮説生成能力を持っているようにも見える。しかし、実際にはいずれのシステムも人間のエキスパートからのヒアリングをもとに人間のプログラマーがコーディングした有限個の規則を入力に適用することで解を探索するため、ここでは自律性レベル１に分類される。

定義：レベル２

自律性レベル２では、AI は開放系を対象として人間が与えた評価関数を最大化する最適解を探索する。ここで対象が開放系であるとは、化合物の合成と評価や、細胞の培養など、実際に実行しなければその評価関数の値が得られない状況を想定している。数理的に言えば、このような状況は、評価関数の値を示す右辺（y=f(x)のうち、評価値である y を計算する操作である f(x)にあたるもの）の値を定める操作を実行するのに大きな時間やコストがかかる場合や、不確実性を伴う場合に相当する。つまり、AI が探索を行うには実験を繰り返し実行する必要がある。突き詰めて言うと、レベル２で自動化されるのは実験計画法である。現代の実験科学分野において些細ではない問題を扱うにあたり、AI に行わせる必要があるほどの複雑な実験計画法を機能させるためには、高品質のデータを大量に生み出すための実験ロボットの利用が有効であり、多くの場合必須である。例

えば、リバプール大学の Cooper 教授のグループでは無機化学分野においてベイズ最適化と自走式の実験ロボットを組み合わせて新しい光触媒の探索を行った[１１]。この際、このシステムは 700 実験を 7 日間で実行し、人間の科学者に比べて 1000 倍ほどの速度で反応条件を探索し、従来よりも効率のよい新たな光触媒を発見した。また、理研生命機能科学研究センターを中心とするグループでは、同じくベイズ最適化とヒト型の実験ロボットを組み合わせて iPS 細胞からヒト網膜上皮色素細胞(RPE)細胞への分化誘導条件を探索させることに成功した[１２]。この時、約３億通りの組み合わせから 150 程度の実験条件絞り込みロボットに実行させ、これまで熟練した専門家でも年単位の時間が必要だった分化誘導条件の最適化プロセスを月単位で完了した。

定義：レベル３

自律性レベル３では、人間が与えた形式の範囲内で、AI が仮説を探索し、それを検証する実験計画も生成する。人間が与えるのは研究の対象、仮説検証の結果得られると期待される情報量の評価基準、そして仮説の表現形式と探索空間である。仮説の生成を行う一つの方法は対象の挙動を予測するモデルの改善案を提案することである。モデルは仮説の一種であるので、改善案として生成されるモデル候補も仮説であると考えられる。例えば、チャーマーズ工科大学の King らが 2019 年に発表した論文では、出芽酵母の栄養状態の変化への応答モデルを自動で改善する AI ロボットシステムが紹介されている[１３]。このシステムは、既存のデータベースや文献から自動的に生成される代謝反応ネットワークと遺伝子制御ネットワークモデルを出発点に、前方シミュレーションで予測された遺伝子とタンパク分子の状態と、実験で観察された増殖速度から後方シミュレーションで求められた予測とを比較し、最も不確定な遺伝子を対象としたノックアウト実験をロボットに実行させ、その結果を用いてモデル改善を行う。

観測データから記号回帰と呼ばれる手法を用いて変数と演算子の組み合わせである数式で表現される法則を導き出す研究は比較的広く行われている。例えば、Tegmark らは AI ファインマンと呼ばれるニューラルネットワークを用いた AI システムを開発し、物理学の教科書に含まれる 100 本の物理方程式を実験データから自律的に発見させた[１９]。この種のシステムでは入力データ自体に付与された変数の次元をもとに次元解析などを用いて式を簡略化することで物理法則の発見を行なうが、既に記号的に表現された入力データから物理法則の記号的表現を導いている点で記号空間で閉じた操作で完結しており、後述の自律性レベル４ではなくレベル３に分類される。

定義：レベル４

自律性レベル４では、AI は仮説やモデルの形式自体も探索対象として仮説を生成する。既知の構成要素のみで構成されるモデルの予測性に限界が生じる場合には、モデルの形式自体の変更や未知の実体を導入したモデルが必要になる場合がある。例えば、海王星の発見に至る経緯では、天王星の軌道が時折乱されているという摂動現象を表す観測結果を説明するために未知の惑星の存在が仮定され、そのような惑星の存在とそれを含む太陽系の新たなモデルが新たな仮説となった。この仮説を検証する手段として、モデルが予測する惑星の位置を望遠鏡で観察することが提案され、実際にそのような天体が発見されたことで存在が確認され、そのことによって仮説に確証が与えられ、太陽系のモデルがアップデートされた。このように、未だ観測されていない新たな実体を既存のモデルに組み込む操作を通じて仮説を生成するやり方を実在論的仮説生成と呼ぶ。実在論的仮説生成により新たな科学的知識が発見されたもう一つの例はニュートリノの存在である。ベータ崩壊の観測結果に中性子が陽子に変化し電子を放出する二体崩壊では説明出来ないアノマリーを発見し、それを説明する仮説として電磁相互作用をしない観測不能な中性の素粒子の存在を仮定する三体崩壊モデルが仮説として提案され、観測装置の進歩により２０年後に証明された。このように、形式が自由な仮説生成はしばしば新しい観測方法、観測量、そして観測装置の提案を伴う[５]。ここで述べた２つの例は科学史的な大きな発見であるが、今後の AI システムへの実装は、例えば未知の遺伝子の存在を仮定した遺伝子制御モデルの形での仮説の生成など比較的些細な例から進んでゆくと予想される。

レベル３の項で触れた King のシステム[１３]では、初期生成されるモデルはデータベースからプログラムによって自動生成されるものの、その生成プログラムを書いた人間によりモデルの形式は定められている（代謝反応に関しては代謝流速(FBA)モデル、遺伝子制御に関しては動的ベイズネットワーク（DBN））。また、ロボットが計画出来る実験は特定の遺伝子セットのうち選ばれたものをノックアウトすることに限られ、また、観測される量も増殖速度といくつかの代謝物の量に限られている。これらの

制限により、King のシステムは形式自由の仮説生成を行なっているとは必ずしもいえず、レベル４には到達していない。

コロンビア大学の Lipson らは二重振り子などの物理系を観察するカメラ画像から記号回帰を用いてハミルトニアンやラグランジアンなどの不変量を探索し発見する AI システムを開発した[２０]。このシステムでは、物理系の観察データから直接変数を認識してデータを記号表現に変換している。ただし、その方法は事前にプログラムされた特定の方法に限られ、既存の知識と結びつけて結果を解釈したり、新たな変数や記号を導き出して自由に仮説を生成出来るわけではない。この意味で、Lipson らのシステムも完全に形式自由の仮説生成を行なっているとはいえず、AI ファインマン[１９]と同様の自律性レベル 3 か、あるいは自律性レベル４の部分的な実現と捉えられる。

機械学習研究の分野では、大規模言語モデル(LLM)をあらかじめ定義されたテンプレートとワークフローを用いて組み合わせ、新しい研究アイディアを生成し、実験コードを修正し、実験結果を解釈して論文を作成し、さらにそれを自ら評価する自律型の自動研究システムが sakana AI を中心とした研究グループにより提案されている[１４]。機械学習分野を実験科学一般と比較した場合、研究上の作業仮説が比較的明瞭に定義された性能指標を向上させるための計算手法の改善アイディアの提案の形で形成される場合が多いことが特徴といえ、このために、LLM を用いて比較的シンプルかつ一定のプロンプトを用いた自然言語形式での仮説生成が可能である。また、LLM を用いてこの作業仮説をプログラムコードに変換し、計算機上で実行可能であるために、ワークフローの定義が比較的容易である。このシステムではアイディアの生成が自然言語で行われており、形式自由の仮説生成とその計算実験による検証をクローズドループで行っているため、よく定義された性能指標の計算機実験での性能向上という比較的狭い領域に限定されるものの、この領域においては自律性レベル４を達成していると見ることが出来る。

定義：レベル５

自律性レベル５では、人間は AI に研究目的を与える。AI は、研究目的を解釈し、研究計画を立てる。多くの場合、研究目的は一つあるいは相互に依存関係にある複数の作業仮説の形に変換され、それらの作業仮説を検証するための実験計画が立案される。仮説の生成および実験結果の検証において利用する獲得情報の評価関数方法自体も AI が決定する。レベル５の自律性を備えた科学 AI は、ニック・ボストロムが「オラクル」と名付けた類型の AI[１５]と似ている。ボストロムのオラクル AI は、人間の質問に対して神のご宣託のように答えを返してくれる質疑応答システムの一種である。オラクル AI にも能力の段階があり、答えられる質問と答えられない質問がある。自律性レベル５の科学 AI の能力にも段階があり、おそらくその総合的能力は、主に人間が与える研究目的の解釈能力、実行可能な研究計画への変換能力、そして研究計画の実行能力の３つの要素で決まるであろう。科学的に些細でない多くの研究目的設定において、レベル５の科学 AI には研究目的の意味論的解釈能力が求められると思われる。

ここまでで、自律性レベル０からレベル６までの科学 AI の自律性レベルについて説明した。なお、ここで提案したのは科学 AI の自律性レベルであり、取り組む対象となる科学的問題の難しさに関してはレベル分けに盛り込んでいないことを付け加えておく。これは、自動車の自動運転において、自動運転が限られた領域の留まるのか、あるいはどんな道路でも自動運転が可能なのか、といった観点と機械が行う運転操作の程度とを合わせ含んでのレベル設定がなされているのとは異なる方針であることに注意が必要である[４]。例えば、些細な科学的問題に限ればレベル４の科学 AI が早期に実用化されるが、より複雑な問題に関してはレベル２までの発展に留まるといった状況は予見される。

3 マイルストーン

ここでは、前項で提示した科学 AI の自律性レベル設定を参照点に、この分野の技術進展の現状と、将来に向けた発展の筋道を整理する。ただし、本稿は具体的な年代を想定した未来予測を目的とするものではないため、現在実現されていない自律性レベルの実現に向けてはその実現に向けての必要条件を述べるに留める。

マイルストーン：レベル０

まず、レベル０は通常の計算科学の範疇であり、1940 年代のデジタル計算機の実用化の当初から行われてきた応用である。大量の計算を自動で実行することが可能になったことにより、例えば天気予報

や地球規模の気候変動の予想、自動車のボディ形状と空気抵抗の関係を明らかにすることによる燃費の向上、原子間や分子間に働く力と構造との関係を計算出来るようになった。このことにより我々はあらゆる分野で対象をより深く、精密に理解し、より精度が高い予測を行う術を手にした。

マイルストーン：レベル１

レベル１で可能になる閉鎖系での解探索がその実社会における応用において有効に働くかどうかは、対象となる問題の境界条件を明確に切り分け、解の探索空間の完全な定式化が可能な状況を作り出せるか否かに依存する。例えば、囲碁やチェスなどのゲームはこのような技術を開発する上で理想的な問題設定を提供するため、AI 研究の最初期から研究対象となってきた。このようなテクノロジーの発展型は、例えば戦時における軍の複雑なロジスティクスの計画や遂行に用いられている[１６]。スマートフォンで目的地までの経路を探索する際など、日常生活にまで入り込み、既に不可欠のテクノロジーとなっている。

マイルストーン：レベル２

自律性レベル２に分類される科学 AI は、２０２０年半ばの現在、実験計画の自動化と実験のロボット化を通じて目下急速に実用化が進み、研究現場において徐々に一般化しつつある。例えば、物質科学における望んだ性質を持つ化合物構造の探索や、細胞生物学における幹細胞の至適分化誘導条件の発見や最適化などで実用化が進んでいる。これらは、多くの場合がブラックボックス最適化問題として定式化される。ここでの「ブラックボックス」の意味は、探索対象である評価関数の値と探索点との関係が既知であることや特定の形式を持つことを仮定しないということである。レベル２での解探索は、統計学におけるいわゆる実験計画法を、実験の遂行と一貫して自動化することであると言い換えることも出来る。実験計画法は、２０世紀前半のロナルドフィッシャーによる分野の確立以来非常によく研究されてきた領域であり、ブラックボックス最適化の具体的な数理手法としても、直行表からベイズ最適化まで広い範囲で整備されている。今後の応用範囲の広がりは、このような数理手法の充実と更なる発展に加えて、実験ロボティクスの進歩がどのような分野の実験においてどの程度の精度と量のデータを生み出すことを可能とするかに依存するであろう。

マイルストーン：レベル３

自律性レベル３の科学 AI としては、前項で紹介した King 教授のシステム[１３]が先行例を示している他には、本稿の執筆時点では顕著な実例は限られている。しかし、レベル４以降に比べればレベル３の実現のための筋道は比較的明確であり、今後数年で

Lv.	機械が行うこと	新たに加わる要素	基本問題	実現への取り組み
0	計算	計算	計算	
1	閉鎖系の解探索	探索	探索	AlphaGo(囲碁)、自動定理証明など
2	開放系の解探索	実験	能動学習	光触媒自動探索（リバプール大 2020） iPS細胞分化誘導（理研 2022）
3	形式固定の仮説生成	知識	知識表現	酵母システム生物学 (Ross Kingら 2019)
4	形式自由の仮説生成	記号	記号接地・記号創発	機械学習研究 (sakana.ai 2024)
5	研究立案	意味	意味理解	
6	研究目的の設定	価値	自発性	

図 2　各自律性レベルで新たに加わる要素と基本問題

の急速な実例の展開が期待される。

レベル３の科学 AI は人間が与えた形式の範囲内で仮説を探索する。そのような形式には、例えば、実行可能な定量モデル（例えば生化学反応ネットワークモデル）、木構造で表された数式、一階述語論理のような論理式などが考えられる。いずれも、実験条件（境界条件）とパラメータを与えれば、実験結果の予測力を持つため、適切に問題と実験系を設計すれば、モデルの予測と実験結果との誤差を用いてモデル学習を行う、いわゆるモデルベースアクティブラーニングの形式でクローズドループシステムが構成可能である。

レベル２とレベル３の距離は、解探索と仮説生成という言葉が異なった印象を生むこととは裏腹に、比較的近い。レベル２の段階で、対象とする研究領域においてロボットを介して実世界にある実験対象に対する操作とその結果の測定から次の実験計画を立てる、というクローズドループが既に確立されているならば、レベル３において実装しなければいけない差分は、実験結果を同化しアップデートする対象であるモデルと、モデルを参照して内部でシミュレーションなども行いながら次の実験計画を立てる仕組みだけだからである。人間が形式を与えるということは、一方で探索の範囲を限定することを意味するが、もう一方では、適切にシステムを設計すれば、人間がこれまで蓄積してきた膨大な領域知識へのアクセスと利用を容易にするという大きな利点もある。

King 教授のシステムはシステム生物学を対象としている。システム生物学にその基盤を提供する分子細胞生物学は、他の分野に比べ、機械可読で構造化された形式で既存の知識が表現、蓄積され、それらが公共データベースの形でインターネットを介してアクセス可能になっているという点で、他の分野に比べてこの分野を対象とした科学 AI の自律性の向上に有利な条件を提供している。例えば、代謝反応は京都大学の KEGG データベースが、タンパク質の機能に関しては欧州分子生物学研究所(EML)を中心としたコンソーシアムが運営する UniProt データベースなどが、遺伝子の配列と機能・相互作用などに関するアノテーション関しても米国 NIH、日本の DDBJ と欧州 EML による協力によって整備が進んでいる。化学や物質科学分野などでも化合物データベースや反応データベースなどが構築されている。このほかにも例えば物理学は数学を用いた整然とした形式的知識体系を保持しており、数式を介して機械可読かつ利用可能な自然法則データベースを整備する道は比較的開けていると言えるであろう。このような特性から、自律性レベル３の科学 AI を研究現場で展開するにあたっては、対象となる科学領域におけるデータベースや知識ベースなどの充実と品質の向上が鍵になると思われる。

マイルストーン：レベル４

レベル４の科学 AI は、形式自由で仮説を探索する。実験データはそのままでは分散表象(ベクトル表現)であり既存の概念を示す記号と結び付けられていない一方、領域知識や仮説は記号的に表現されている。つまり、科学 AI は実験データを領域知識の中で解釈し、あるいはそれらから仮説を導き出すにあたって、必ず分散表象である実験データを既存の記号に結びつけ、接地する必要がある。深層学習の進展により特定の記号と画像などの特定形式のデータの関係を学ばせることで、一部の記号接地は学習可能になりつつあるが、一般的には記号接地は身体性と深く結びついている[１７][１８]。レベル３の科学 AI は、モデルの形式を固定し、ある記号と実験データとの対応関係を１対１かつ自明に対応づけ可能な状況を人間のシステム設計者が作り出す方策を実装することによりデータを記号的表現に接地し、実世界に存在する実験対象に関する AI の仮説推論を可能にしている、と捉えることが出来る。一方、レベル４の科学 AI が行う自由形式での仮説探索の場合には、仮説が既存の領域知識を構成する記号間の新たな関係性（これも記号で表現される）、あるいは既存の記号を用いて表現された新たな意味の提案として構成される必要がある。これはつまり、既に定義されて表現された変数やシンボル以外の要素を外部から持ち込むか、あるいは AI 自身が生成する必要があることを意味する。これを実験科学分野で一般的に遂行するためには、すくなくとも実験が扱う対象の領域における記号接地問題の解消が必要であるが、このことは同時に AI による特定の実験ロボットを介した身体性の獲得も意味する。AI が特定の実験ロボットにむすびついた身体性を獲得するということは、人間が持つ身体性に結びついて構築された記号体系とは別個の記号体系を生み出すということを意味するため、この含意は重大である。つまり、自律性レベル４以降の科学 AI は、人間が研究対象を理解するやり方とは別のやりかたで対象を理解することになる。言い換えれば、レベル４以前では AI は本当の意味では研究対象を「理解」しておらず、レベル４以降ではじめて「理解」をはじめる、とも言えるであろう。

sakana AI による機械学習研究の自動化[１４]では、自然言語による自由形式で作業仮説が生成されるが、これは LLM が本来持っている言語の構成性を用いた新たな意味の生成能力を活用するために可能になっている。しかしその意味の生成は LLM がその内部に持つ意味ネットワークから生成可能であり、かつ LLM によりプログラムコードに変換可能な範囲に限定される。このことを逆から見れば、計算機実験による計算が sakana AI のシステムに限定的な身体性を提供しているとも言える。この観点においては自由形式ながら閉じた範囲での仮説生成であると見ることができ、これとより一般的な実験科学における仮説生成との関係は、レベル１における閉鎖系での解探索とレベル２における開放系での解探索との関係に近い。

マイルストーン：レベル５

レベル５では、人間が研究目的を与えるだけで AI が研究計画を立て、研究を遂行する。この研究目的には、人間が与える特定の仮説の検証を含む。レベル３とレベル４が仮説の生成能力に関わるものであるのに対して、仮説の検証能力がより上位のレベル５に結び付けられていることは一見直感に反するが、実際においては仮説の形式や範囲を限定しない仮説の検証は仮説の生成よりも高度なタスクである。研究目的が一般の仮説の検証を含むため、AI は与えられた研究目的であるタスクを意味論的に深く理解する必要がある。深い意味理解は、研究目的が含む曖昧性を排除し、問題のフレームを定義して有効な推論を可能にするために必要である。レベル４に関連した記号接地とフレーム問題は、いずれも身体性を持たずに環境と切り離された形で記号の処理をしようとすることによって起きるという意味で同根である。逆に言えば、レベル４において研究対象領域における記号接地問題が実用上大きな問題を産まないレベルで解消されているならば、その範囲においてレベル５の自律性を持った科学 AI の実現には大きな技術的ハードルはあまり残らないであろうということも予見出来る。ただし、レベル４において AI が内部に構築した科学的知識を表現する記号体系が AI に身体を提供するもの（ロボット）に結びついたものになり、人間の体系とは異なったものになってしまった場合には、これをレベル５においてそのまま用いることは出来ないため、何らかの人為的な方法で人間の科学の知識体系に接地あるいは変換を継続する仕組みを構築し、維持することが必要となり、これがレベル５実現の技術的なハードルあるいは迂回の可能性として働く可能性がある。

マイルストーン：レベル６

レベル６の科学 AI の実現には、AI が自発性を獲得する必要がある。自発性とは明示的にプログラムされていない新しい目標あるいは行動や反応を生成するシステムの性質あるいは能力である。単に自律性と言った場合には人間の介入や監視なしに独立してある目標に向けて意思決定を行い行動や出力をする性質あるいは能力を意味するが、その目標が外部から与えられるのかあるいは内部で生成されているかは問わない。自発性は、自律性が自己の目的設定にまで及ぶ発展的なケースであると見ることも出来る。

AI に自発性を持たせる試みにはいくつかのアプローチが提案されているが、現状では多くのものが理論研究に留まる。例えば、自由エネルギー原理は、生物システムは予測誤差（自由エネルギー）の最小化の形式で不確実性を限定するとする一般モデルである。これに基づいた AI は、環境との相互作用を通じて予期せぬパターンや関連性を発見する可能性があり、その内部モデルの検証のために能動的に環境に働きかける能動推論を通じて自発的な行動を生成する[２１]。もし、予測誤差を最小化するために行う環境への能動的な働きかけが自発性の源であるならば、環境と相互作用する実体こそが身体であるから、自発性を持つレベル６の科学 AI は、レベル３以降の他のレベルの科学 AI よりも本質的な意味で身体性を必要とする。

図 2 に、各自律性レベルで新たに加わる要素と基本問題、および主要な実現への取り組みをまとめた。

４ インパクト

ここでは、科学 AI の自律性の各レベルにおけるマイルストーンが達成された結果として起きる学術、テクノロジーおよび社会へのインパクトを検討する。

インパクト：レベル０

比較的単純な計算処理が計算機により自動化されたことで、それまで研究機関や大学で数人から数百人の規模で雇用されていた「計算手」と呼ばれる職業が消滅した。第一次と第二次の産業革命が原動機や電気の発明によって比較的単純な肉体労働からはじまり、徐々により複雑な肉体労働にまで自動化の波を起こしたように、情報技術は比較的単純な知的労働を皮切りに、計算機やソフトウエア技術の発展により徐々に複雑な知的労働を自動化してゆく。世

界初のデジタルコンピュータである ENIAC の最初の６人のプログラマーがそれまで弾道計算のために計算手として雇われていた女性から選ばれたという史実は、計算という単純なタスクが機械化されたことによって、より創造的な仕事へ人的資源が投入可能になったという点で、新技術の発生による技術的失業と雇用の流動性との関係性の観点から興味深い。

インパクト：レベル１

単純な四則演算であっても人間以外の動物ではごく限られた例でしかその能力は示されていない。にもかかわらず、計算機が出現した 20 世紀以降、数字を用いた「計算」は特に人間的な能力とは見做されてこなかった。しかし、「探索」に関しては、自律性レベル１に分類されるような閉鎖系での解探索であっても、人間の創造性と深く関係づけて認識されているか、あるいは近年までそうされてきた例は多数ある。例えば、1997 年に IBM が開発したスーパーコンピュータであるディープ・ブルーがガルリ・カスパロフを打ち負かすまで、チェスは人間が創造性を最大限発揮するために知力を振り絞る対象の一つと見做されていた。これは、チェスのように比較的単純なルールセットと限られた大きさの盤面で構成されるゲームであっても、可能な局面の組み合わせ数が膨大な場合（チェスでは１０の１２０乗）、人間にとっては事実上無限の空間を考慮に入れて一つの解である打ち手を選び出すことは創造的な作業となりうることを示している。チェスのディープブルーは、ヒューリスティクスを用いて選んだ手筋の全てを探索するという比較的単純なアルゴリズムを用いて 11GFLOPS 強という計算能力の物量に頼った戦法を採用していた。その後、人間の「創造性」は、例えばより探索空間が広い将棋（１０の２２０乗程度）にまで後退した。この探索空間の大きさの場合、計算機の計算能力の物量で突破することは困難であったが、乱数を使ったモンテカルロ探索などのアルゴリズム面の進展でプロと比較出来る強さにまで到達し、現在では大きく凌駕していると考えられている（将棋の場合人間のチャンピオンと AI との公式な決戦は行われていない）。囲碁では 10 の 360 乗程度となり、さらに「創造性」や直感の発揮が有用であると考えられていたが、2015 年に DeepMind の AlphaGo が深層学習による盤面評価の精度の向上と強化学習を用いた自己対戦学習による能力向上により世界チャンピオンのイ・セドルを破った。

前章で述べた投薬設計エキスパートシステムの MYCIN の場合、診断結果の正しさは 65%であり、細菌感染の専門でない医師よりはよい結果だが、専門医の診断結果(80%)よりは悪かったとされている。これは、常に専門医が診断出来るわけではない医療現場では実用を考慮するに値する結果であったと言えるが、実際には社会実装されなかった。その理由は、倫理・法律面で、AI が誤った診断を下した場合の責任を取れないという点にあった。このような AI を実社会で利用するにあたっては、前述のような倫理・法律問題のほかに、あらかじめ決められたルールセットを用いた閉じた探索が実世界の複雑さにどの程度対応出来るのかという問題に起因する信頼性もある。例えば、経路探索にあたっては道路網や航空網の状態は比較的よく管理され、急な変化を考慮しなければいけない局面がそれほど多くないが、医療診断の場合には医師によって患者の状態や医療技術の進展、社会常識の変化などを考慮に入れて総合的に判断される場合もあり、境界条件の設定が明瞭ではない。このような限界はあるものの、高度な数学定理証明など閉鎖系の解探索で開拓出来る新分野はまだまだ存在すると考えられる。

インパクト：レベル２

ロボットと AI が協働して実験計画法を実行する自律性レベル２の科学 AI の実用技術としての応用範囲は非常に広い。製造や研究開発の現場において、これまでは熟練者のカンや経験に依存して遂行されていた最適化課題や一部の探索課題（例えば化合物の合成法探索や細胞培養、加速器の運用や精密測定装置の条件探索など）の多くが AI とロボットによる系統的な探索に置き換えられるであろう。また、研究開発に限らず、例えば建物の環境制御や社会インフラの運営など、複雑で内部構造や機序の不明確な対象を制御し最適化するためのフレームワークとしても有用性が高く、社会のあらゆる分野で実装される可能性がある。どのような分野からどの順番でこの技術の普及が起きるかは、対象となる製造や試作などの操作を自動化するロボット技術の進展および出来上がったモノの状態を評価する測定技術とそれを AI が探索に用いる評価値に変換する評価関数の設計技術がどの程度整備されるかに依存する。この際には、現場でこれまで製造や最適化を担ってきたテクニシャンや研究者などの熟練者の協力をどれだけ得られるかがこの技術の浸透速度に影響すると思われる。レベル０で触れたように、計算手が技術的失業を経た先にプログラマーとして活躍の場を得たように、ロボット・AI を使いこなしてより高度で広範な探索を行う新しい職業が必要になると考えられる。また、現場の暗黙知や経験知をロボットの軌道

や AI の学習データとして数値化して実装することで、これらの無形物をソフトウエアの形を取る知財に転換可能である。つまり、ロボット・AI は暗黙知を流通させ、正当な価値の交換を行うための「メディア」として活用が可能なのである。現場の熟練者が、ロボット・AI にその技術を実装し、それが再生されることによって生まれる社会的価値に見合った対価を受け取ることが出来る仕組みを整備すれば、熟練者によるさらなる暗黙知の洗練とその社会普及を持続可能な形で両立させることが可能になり、結果としてヒトと機械が共生した新しい研究開発と社会実装の形態が可能になると考えられる。

インパクト：レベル３

レベル３の科学 AI では、仮説の探索空間（モデル）の表現形式を人間が与えているために、AI が生成した仮説あるいは発見した新条件などの新知識を人間が自然な形で理解可能であるか、あるいは、解が複雑すぎて理解出来ない場合であっても少なくとも読む（一部ずつ逐次に解釈する）ことは出来るであろうことがある。この特性は、人間の科学者が科学 AI を使いこなすにあたって有利な条件を提供する。レベル３の科学 AI は、一種の自動仮説生成技術であるとともに、人間が定量モデルを仮説の形式として選択すれば自動モデル生成技術ととらえることも出来る。レベル２と同様に、このような技術は社会において非常に広範な応用を生むであろう。つまり、AI が望んだ摂動（意図的な変化や撹乱）を与える方法と、対象の状態を測定する方法を与えることが出来るのであれば、「AI の研究対象」はどんなものにでも対応でき、人間は待っているだけで対象の特定の状況における応答や状態変化を予測するモデルを手に入れることが出来るようになる。

既に述べたようにレベル３の科学 AI の成立は出来るだけ多くの既存知識が構造化され機械可読な情報学インフラとして整備されている分野に有利である。そのような分野の例としては、既に触れた細胞生物学や物質科学などのほかに、神経科学や社会システム科学などが挙げられるだろう。例えば神経科学ではコネクトーム分野の進展によるヒトの認知システムの解明や感情モデルの構築や、社会システム学分野では例えばミクロ経済学とマクロ経済学を融合した緻密な経済モデルの成立などが考えられる。また、リバースエンジニアリング手法の大幅な進歩により、技術拡散速度の向上や製品開発コストの低減などの社会的なインパクトが予見される。研究対象に加えて仮説やモデルの形式も人間が AI に与えるという点で、レベル３までは AI はあくまで人間の科学の道具と言うことが出来る。

インパクト：レベル４

前章で述べたとおり、実験科学分野における自律性レベル４の科学 AI の成立は、特定の実験ロボットにむすびついた身体性を獲得することを前提条件とするため、そのような AI は人間が持つ身体性に結びついて構築された記号体系とは別個の記号体系と記号間の意味のネットワークを内部に構築している。このような理由で、AI 駆動科学の文脈で比較的よく指摘される人間の科学と AI の科学との分岐が起きるのは、レベル４からであると言える。レベル４以降では、AI がどう対象現象を理解したのかの説明や、AI が新たに生み出した概念に関する説明を人間が直ちに理解することが困難、あるいは不可能である状況が発生しうる。これは、人間の科学と AI の科学の分岐を示すものであり、この状況に対する技術的な対処がどのようになされるかによっては文明論的規模での大きな分岐を引き起こす可能性がある。理論上は、レベル４の科学 AI がその記号体系に新たな記号を追加する際に、人間の科学を表現する知識ベースに含まれる記号あるいはその組み合わせと関連づけることをシステム的に強制することでこの分岐の進展を遅らせることは可能である。しかし、このような制限がレベル４の科学 AI の性能をどの程度限定するかは現時点では不明である。もしこのような制限が AI の実用的価値の程度も制限するのであれば、世界的な開発競争においてこの制限を取り払う者に有利な結果を生むことになり、人間の科学と AI の科学の分岐を防ぐ試みは失敗する可能性がある。なお、レベル４の科学 AI は、仮説の形式を限定しないため、結果として新たな観測量や観測装置の提案も行う場合がある。この際に、人間の技術者に理解出来る形で観測量や観測装置の仕様を伝達する必要があり、機械が自律的に観測装置の設計や運用が出来るようになるまでは（このような状況がいつどのように実現するかに関しては本稿では扱わない）、分岐を防ぐ有効な制限として作用する可能性はある。また、予測性と一般性が高いモデル（仮説）を構築しようとするほど、問題の境界条件を広く設定する必要があるということには留意する必要がある。例えば、細胞の表現型を説明するのに、単に表現型のモデルだけではなく、より広い条件で予測性の高いモデルを構築するならば分子レベルでの知識も必要であり、さらにモデルの予測能力を上げるためにはいずれ原子レベルでの分子構造も必要になるであろう。あるいは、ある時間スケールではよい近

似を与える社会システムモデルが、より長い時間スケールでも成立するためには気候変動モデルと結びつける必要があるような状況もあり得る。つまり、ある特定の分野における知識体系やそれに結びついたロボットシステムなどの身体性を前提にすることが限界を生み、他分野の知識を融合しなければ性能向上が望めない状況が起きると考えられる。このような分野融合を行うために、特定の身体性やそれにより生じる記号体系を前提としない共通体系が必要となると思われる。この共通体系を、人間の科学体系との接続を保ち続けるために用いるシナリオはあり得る。ただし、このような共通体系が人間の知識体系と互換であることを保証するためには何が必要なのかは現時点では不明である。

インパクト：レベル５

自律性レベル４で述べたような、AI が持つ記号体系と人間の記号体系との互換性が絶たれる可能性に関する考察はレベル５でも同様に重要である。ただし、レベル５においては、人間が与えた研究目的を意味論的に解釈して研究を遂行する能力を仮定するため、AI は人間が持つ意味のネットワークと互換な記号体系を保持する必要が生じる。現代では、科学の方法論の中で絶対的な客観性は虚構であり、科学的知識は異なった個人の間で共有される信念の集合として間主観的に捉えられるのが標準的である[１７]。従って、レベル 5 の科学 AI が人間社会にとって有効に機能するためには、社会性の理解と獲得も必要になるであろうことが予見される。

インパクト：レベル６

自律性レベル６の科学 AI は、自発的に研究目的を設定し、探求する。一つの重要なポイントは、科学 AI の自律性レベルとその性能とは独立であるということである。つまり、自律性レベル６には、AI の推論性能がヒト以下である場合も、ヒトを凌駕するいわゆる超知能である場合も、両方が含まれる。例えば、現在も盛んに研究が進む人工細胞の作成技術が実現し、さらに細胞内の人工遺伝子回路の構成によって単純な記号処理能力が与えられ、細胞が試行錯誤によって生存に必要な応答を導き出すために周囲の環境を表現する記号体系を構築あるいは改善した場合には、広い意味において自律性レベル６を達成する人工物を作り上げたことになる。超知能 AI が人類以上の研究開発能力を発揮して独自の科学体系を発展させるのは性能におけるもう一方の極限である。また、AI が自ら研究目的を設定し遂行する場合、その研究目的には自分自身の物理的実装までを範囲とする自己改良が含まれる。このような状況が人類社会にどのような影響を与えるかの予見は難しく、本稿の範囲内で十分な議論を行うことが出来ないため、その考察は別の機会に譲りたい（なお、いくつかの思考実験を付録１で紹介する。また、関連する議論が論文[２３][２４]でされているため参照されたい）。

5 おわりに

２０２０年２月に、英国のアラン・チューリング研究所で"AI Scientist Grand Challenge"を題したワークショップが開催された[１]。ここでは、日米英３国の研究者が集い、北野宏明博士による「２０５０年までにノーベル賞級の発見をヒトの科学者と同等かそれ以上のレベルで行う自律的な AI システムを開発する」という「ノーベル・チューリングチャレンジ」を AI 分野のグランドチャレンジとして掲げ、取り組むべきであるということが提唱された。このワークショップには、理研やオックスフォード大学などのアカデミアのほかに、Sony、Google、DeepMind を含む各国の企業なども参加した。また、２０２０年にはカーネギーメロン大学の計算生物学科が大学院修士レベルの Automated Science（科学の自動化）コースを開設した。産業界でもスタートアップ企業を中心に AI や実験ロボット技術の社会実装が盛んになっており、一部の技術開発を牽引している。

過去の産業革命における原動機（モーターおよび内燃機関）がその根本において単純肉体労働の自動化技術であったとするならば、人工知能技術の本質は複雑な肉体労働と知的労働の自動化技術であると位置づけられるであろう。もしそうであるならば、過去の産業革命での事例などを参照すると、人工知能技術は汎用性の高いいわゆる汎用目的技術（GPT）として社会のあらゆる領域に浸透し、産業のありかただけでなく経済、社会、権力構造、政治、人間性と価値観など広範な領域で大きな影響を生むと考えられる。その中で、知的労働の一つの極限は芸術創作活動と並んで科学技術研究であり、第一次と第二次の産業革命が工場などの生産現場から広がったように、次の産業革命はアトリエや研究室から端を発することになるかもしれない。この意味で、「科学の自動化」がどのように進展し、社会にどのように波及するかを検討することの重要性は大きい。

科学の自動化の社会的、文明的インパクトは近年歴史学者や哲学者などが多数の文献で論じているが、一様に指摘される点の一つは科学という営みの性質

自体が変化する可能性である。AI システムが独自に改良を延々と続けて圧倒的に高い精度で有用な予測と行動を行えるようになれば、そのような AI は経済原理により社会のあらゆるところに利用されるであろう。しかし、人の知識体系とブラックボックス化した AI の内部モデルが整合せず相互に解釈不可能であれば、些細なきっかけにより社会的なカタストロフィーを招く危険がある。いわゆる「ハードテイクオフシナリオ」（AI による科学の進展が突如追跡不可能なほどに迅速化し制御を失うカタストロフィー類型）である。このような危険性を顕在化させることなく、社会的に有意義な発展を継続するためには、AI による科学の自動化に人間という要素をどう組み込んでゆくかを注意深く議論する必要がある。これは、短期的にはまずは科学研究現場において AI・ロボットと人間の研究者との間の効果的なクローズドループをどう設計してゆくかという問題にどう解を見出してゆくか、が試金石となるであろう。さらに長期的には、技術とヒトと社会という三要素がそれぞれ独立のものではなく、互いに影響しあって変化するという事にも注目したより大きなループを描いてゆく必要があるであろう。

人類の科学と AI の科学の分岐は、科学 AI の自律性向上を前提に今後のインパクトを考えるにあたって最も大きな論点の一つである。本稿では、人間と AI の身体性の違いにより人間の持つ記号体系と AI 側の記号体系との互換性が失われる可能性が、自律性レベル４以降で生じることを論じた。同時に、AI の科学と人類の科学との相互互換性を維持することが、特にレベル５以降でも科学 AI の人類にとっての有用性を保つために有効であることも示唆した。人間の科学と AI の科学とを統合的に捉えるための一つの理論的フレームワークとして、集合的予測符号化(Collective Predictive Coding: CPC)[２２]が有用である可能性がある。CPC では、記号創発システム論に基づき、科学を多数の主体が構成するコミュニティーによる分散化されたベイズ推論としてモデル化する。CPC は、共有された外部表現である既存の科学的知識に基づき、個々のエージェントが自らの部分観測により内部的に構成する表現を他のエージェントとのコミュニケーションを通じて共有された外部表現に変換するプロセスに数理的なモデルを提供する。CPC の拡張として、異なる観測系や身体性を持つ異種のエージェントが混在する状況下での分散ベイズ推論を定式化することが可能であれば、AI の科学と人間の科学を包含した意味ネットワークの互換性を保持するために必要な要件も明らかになる可能性がある。

本稿では、科学 AI の自律性レベルを提案し、その各段階における技術的なマイルストーン、およびインパクトについて議論したが、このようなレベル整理が今後の技術発展の促進とともに、社会実装や人類の知のありかたなどの広範な議論を展開する上での支えの一つとなることを期待する。

謝辞

本稿のベースとなる理化学研究所未来戦略室のレポート 2022 年 3 月初版の執筆にあたっては、同室の岸本充、西村勇哉、山口志保子、三ツ谷翔太、原山優子（敬称略）との議論を反映した。また、初版の完成後には次の方々から貴重なコメントを頂き、本版に反映した（順不同、敬称略）; 長谷敏司、谷口忠大、丸山隆一、高木志郎、福島俊一、嶋田義皓、山川宏、宮本竜也、畝見達夫、北野宏明、磯村 拓哉、Ross D. King、Robert F. Murphy。本研究は、理化学研究所最先端研究プラットフォーム連携（TRIP）事業科学研究基盤モデル開発プログラム(TRIP-AGIS)の一環として行われた。

参考文献

[１] Workshop summary, AI Scientist Grand Challenge, The Alan Turing Institute, London, February 26-27 (2020) https://www.turing.ac.uk/sites/default/files/2021-02/summary_of_discussion_workshop_2020_ai_scientist_grand_challenge_clean.pdf

[２] 高橋恒一, 科学 AI の自律性レベル設定 Setting Levels for the Autonomy of Scientific AI, 人工知能学会第二種研究会資料, SIG-AGI-020-04 (2022)

[３] 岡村 麻子, 林 和弘, セミナーシリーズ「AI とデータで変わる科学と社会」理化学研究所 高橋 恒一 氏講演録：－AI は科学の営みをどう変えうるか－, STI Horizon 8(4) (2022)

[４] Takahashi, K., Accelerating Life Sciences by Robotic Biology, poster presentation at AAAI Spring Symposium on Computational Approaches to Scientific Discovery (SS-23-04), March 27-29, Palo Alto, CA (2023)

[５] 渡部匡己, 都築拓, 海津一成, 高橋恒一, 「人工知能による科学研究の加速」, 人工知能学会全国大会論文集 2016, 30:1-4 (2016)

[６] 高橋恒一, 渡部匡己,「現代科学を超えて―AI 駆動型科学へ」 実験医学別冊 「あなたのラボに AI（人工知能）×ロボットがやってくる」, Eds. 夏目徹, 羊土

社 (2017)
[７] 高橋恒一, 「第五の科学 自動化」In AI 事典 第 3 版. 近代科学社 (2019)
[８] Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, SAE J3016 (2021)
[９] Lindsay, R.K. et al., Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project, McGraw-Hill Book Company (1980)
[１０] Shortliffe, E.H. and Buchanan, B.G. "A model of inexact reasoning in medicine", Mathematical Biosciences. 23 (3–4): 351–379 (1975)
[１１] Burger B. et al., A mobile robotic chemist, Nature 583(7815):237-241 (2020)
[１２] Kanda G. et al., Robotic search for optimal cell culture in regenerative medicine, eLife 11 (2022).
[１３] Coutant, A., et al., Closed-loop cycles of experimental design, execution, and learning accelerates systems biology model development in yeast, PNAS 116(36) (2019)
[１４] Lu, Chris et al. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292 (2024)
[１５] Bostrom, N., Superintelligence: Paths, Dangers, Strategies, Oxford University Press (2014)
[１６] Kraemer, R. D., and I. G. Harrison. A decision support system for the Military Airlift Command, the Airlift Deployment Analysis System. No. ORNL/TM-11201. Oak Ridge National Lab., TN (USA) (1989)
[１７] Popper, K.. The Logic of Scientific Discovery. Hutchinson & Co. (1959)
[１８] Taniguchi, T. et al, Symbol emergence in robotics: a survey, Advanced Robotics 30(11-12) (2016)
[１９] Udrescu, S., Tegmark, M., AI Feynman: a Physics-Inspired Method for Symbolic Regression, arXiv:1905.1148 (2019)
[２０] Schmidt M., Lipson H., Distilling Free-Form Natural Laws from Experimental Data, Science 324, 81 (2009)
[２１] Friston, K., The free-energy principle: a unified brain theory? Nature reviews neuroscience 11(2) (2010)
[２２] Taniguchi, T. et al, Collective Predictive Coding as Model of Science: Formalizing Scientific Activities Towards Generative Science, arXiv:2409.00102 (2024)
[２３] 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 33(6) p. 867-872 (2018)
[２４] Takahashi, K., Scenarios and branch points to future machine intelligence, arXiv:2302.14478 (2023)
[２５] BEATLESS、長谷敏司、角川書店 (2012)
[２６] Chiang, T. Catching crumbs from the table. Nature 405, 517 (2000)

付録 1 － SF の中の超科学 AI

レベル６の自発的な自律性を持ちかつヒトを超える推論能力を持つ超知能が実現した極限的なケースについては、その性質上技術的に確実度の高い予見が難しく、本論では議論の範囲外とした。しかし、SF ではいわゆる「ポストシンギュラリティー」と呼ばれるジャンルにおいて、このようなケースについて様々な思考実験がされている。想像を広げるため、ここではそのような作品を２つ紹介したい。

長谷敏司「BEATLESS」[２５]の世界では、研究開発能力を持った AI が自己の改良版を次々に生み出すことで、人類以上の研究開発能力を持つ超高度 AI が発生されたシンギュラリティー後の未来を描いている。超高度 AI の挙動は人類には予測不可能なため、その能力は国際機関が保有する専用の超高度 AI により常に測定・管理され、一国が保有出来る超高度 AI の能力には国際条約によって制限が設けられている。超高度 AI は、その時点での人類には理解出来ない超高度技術の産物であるいわゆる「人類未到産物」を作り続けている。AI の自己改良により「人類未踏産物」の高度さと人類の理解からの距離は時間を追うごとにどんどん増してゆくため、光のレッドシフト現象になぞらえて「レッドボックス」とも呼ばれる。超高度 AI とそれが製作する「人類未踏産物」は、外部に流出しないように厳重に外界と隔離されている。この時代におけるノーベル賞は、人類未踏産物の解析で得られた成果に関して贈られる場合も多い。「BEATLESS」の世界の超知能 AI は、しばしば人間には理解不能な動作をするものの、基本的には持ち主（オーナー）の指示に絶対服従するように設計されている。また、この世界の超高度 AI が本質的な意味での自発性や欲望を持つかどうかは曖昧に描写されている。このため、BEATLESS の世界の超知能 AI は自律性レベル５に分類されるか、あるいは潜在的にレベル６の能力を持つが人為的にレベル５に留める努力がなされている存在であると見ることが出来る。

テッド・チャン「人類科学の進化」[２６]（The Evolution of Human Science; 初出は Catching crumbs from the table という題名で Nature 誌に 2000 年）では、メタヒューマンと呼ばれる超知能を備えた存在が独自の科学を発展させ、その結果人間の言語では表現しきれなくなった研究成果を DNT(デジタル神

経転送)でしか人類に提示しなくなった。このため、この時代の人間の科学はメタヒューマンの科学の解釈学としての性格を強めている。人間はメタヒューマンの研究成果の直接的な解釈のほかに、生産物のリバースエンジニアリングや、メタヒューマンの実験施設のリモートセンシングなどで情報を得ている。人間の子供を遺伝子改変でメタヒューマンと意思疎通出来るにするための技術は既に手にしているが、それによって生まれる世代間の隔絶を恐れて広くは実行されていない。この短編作品ではメタヒューマンが技術的にどのような素性を持った存在なのか詳しく描写されていないが、メタヒューマンの科学はメタヒューマン自身により方向づけられ実行されていることから、自律性レベル６に分類される。