人工知能・科学・人間のトリロジーの将来

植原　　亮*

要　旨

人工知能（AI）は，科学的探究において非常に大きな役割を担いつつある．しかし，AI は単に科学的探究を効率的に前進させる道具として働くにとどまらず，その使用は科学のあり方そのものに重大な転換をもたらす可能性がある．AI の行う情報処理は，人間には理解困難なブラックボックス的性格がしばしば見られるが，そこからは，AI が科学と人間の従来の結びつきを根本的に変化させ，最終的には人間を科学から切り離してしまうのではないか，といった懸念が生じるのである．本稿では，そうした懸念の内実を明らかにし，それにどのような対応策があるのかを概観・検討したうえで，AI と科学，そして人間の三者が形づくる関係の将来について考察する．具体的には，説明可能・理解可能な AI の開発を目指す研究も対応策として始まっているが，その成功は必ずしも保証されていない．しかしそれでも，たとえば脳の研究と AI の研究とが相補的に進められるような，新たな領域を切り開いていける可能性は十分に残されている．結論部では，これから当面のあいだ続くのは，AI を媒介にした新しい研究領域が次々と登場し，まさにそのことによって AI と科学と人間がいっそう緊密な関係を取り結んでいく将来であるとの見通しを示している．

キーワード：人工知能，AI，科学，哲学，ブラックボックス

The Future of the Trilogy of Artificial Intelligence, Science, and Humans

Ryo UEHARA

Abstract

Artificial intelligence plays a huge role in scientific inquiry. However, AI does not merely act as a tool to advance scientific inquiry efficiently. Its use could lead to a profound shift in science itself. AI's information processing often has a "black box" nature, which is difficult for humans to understand. The concern which arises is that AI will fundamentally alter the traditional link between science and humans and separate humans from science. This paper aims to clarify the nature of such concerns, outline and

*関西大学総合情報学部

examine how to respond to them, and discuss the future of the relationship between AI, science, and humans. Specifically, research has begun to develop explainable and understandable AI as a response. However, its success is only guaranteed sometimes. There remains ample potential to open new areas where, for example, brain research and AI research can be complementary. The concluding section of this paper highlights the prospect of a future in which AI–mediated new research areas will continue to emerge in the foreseeable future. This will bring AI, science, and humans closer.

Keywords: artificial intelligence, AI, science, philosophy, black–box

人工知能（Artificial Intelligence）――以下「AI」と略記――は，科学的探究において非常に大きな役割を担いつつある．しかし，AI は単に科学的探究を効率的に前進させる道具として働くにとどまらず，その使用は科学のあり方そのものに重大な転換をもたらす可能性がある．AI の行う情報処理は，非常に強力ではあっても，人間には理解困難な，つまり人間から見て「ブラックボックス」的な性格がしばしば見られるが，そこからは，AI が科学と人間の従来の結びつきに根本的に変化させ，最終的には人間を科学から切り離してしまうのではないか，といった懸念が生じるのである．本稿では，そうした懸念の内実を明らかにし，それにどのような対応策があるのかを概観・検討したうえで，AI と科学，そして人間の三者が形づくる関係の将来について考察したいと思う．

第 1 節　AI と科学的探究の現在

本節では，考察の出発点として，AI が科学的探究にどのような力を与えているのかを確認しておきたい．そのためにまずは，AI にまつわる用語を簡単に整理し，次いで AI という技術の概要を押さえる．そのあとで，科学的探究で AI が実際に用いられている例をいくつか見ることにしよう．

(1) 用語の整理と技術の概要

AI の研究には，機械による視覚の実現（ヴィジョン）や言語・推論などを中心としたコミュニケーションなど，さまざまな対象を扱う分野がある．その中でも，コンピュータに学習をさせることを目的とした技術的手法の開発を目指す分野のことを「機械学習（Machine Learning）」と呼ぶ（そうして開発された技術的手法そのものも「機械学習」と呼ばれる）．ここでいう「学習」とは，人間が明示的にプログラムすることによってではなく，コンピュータ自身がこの世界についてのデータにもとづいてその性能を向上させることを意味する．機械学習を実現するための学習アルゴリズムとして，サポートベクターマシン，ランダムフォレスト，確率グラフィカルモデル，強化学習などが挙げられるが，本稿で中心的に取り上げる「深層学習」――

「ディープラーニング（Deep Learning）」――もそのひとつである．これらの技術はしばしば組み合わされるし，また深層学習にも手法に応じた下位分類が設けられているが，本稿ではそうした細部に立ち入る必要はない（ただし注１で手短に補足している）．また，叙述を簡潔にするために，以下では「AI」の語で，しばしば限定なしに深層学習を学習アルゴリズムとして用いるものを意味させることにしよう．

AI，とくに深層学習を用いたAIは，与えられたデータにもとづいて学習を行い，さまざまなものを分類，認識，分析，検出，発見，予測する機能を有する．大量の画像データにもとづいて学習を進めさせることにより，新たに与えられた画像に写っているのがパンダなのか，それとも他の対象なのかを分類することができるようになる，というのが具体例のひとつである（図１，cf. 瀧 2020）．学習を済ませたAIに実際にパンダの画像を与えると，それが適当なピクセルに細かく区分けされて入力される．AIはその入力を分析して，「この画像は99.99％の確率でパンダである」とか「猫の可能性は0.01％である」「雄羊の確率はゼロである」といった確率を含んだ分類・認識を行う（つまり出力する）．

では，ここでいうデータにもとづく学習とは何だろうか．深層学習を用いるAIは，動物の脳をシミュレートした構造を有している．脳は膨大な数のニューロン（神経細胞）が互いに結びつくことで構成されている．これを「ニューラルネットワーク」といい，その結びつきの強弱に応じた仕方で，個々のニューロンは他のニューロンから受けとった情報をさらに別のニューロンに伝えていく．図２は，そうしたニューラルネットワークを人工的にシミュレートしたAIの模式図である．この人工ニューラルネットワークでは，そのつどの入力（データ）と出力がうまく対応するように，無数に並べられたニューロン同士の結びつきを強めたり弱めたりすることで，情報をどう伝播するかが更新されていく．大量のデータによりこれを繰り返すことで「訓練」を続けていくと，先の例でいえば，パンダの画像の入力に対し「99.99％で

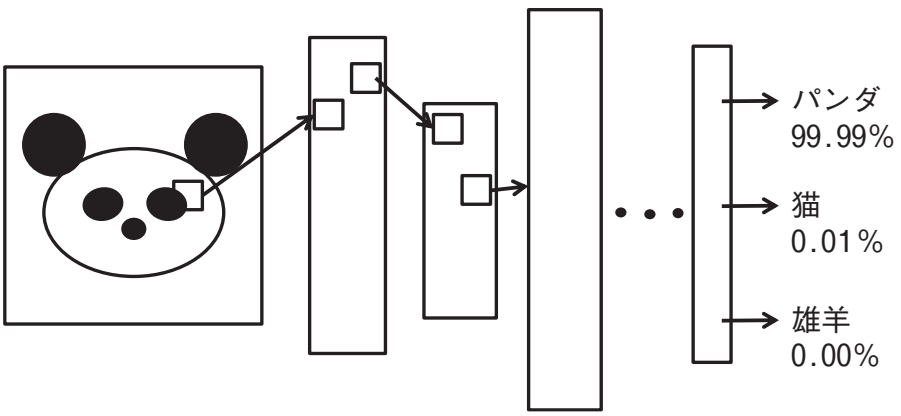

図１　深層学習による画像の予測・分類
（瀧 2020：44の図を参照しながら，ひとつのピクセルにのみ注目した簡略図を作成）

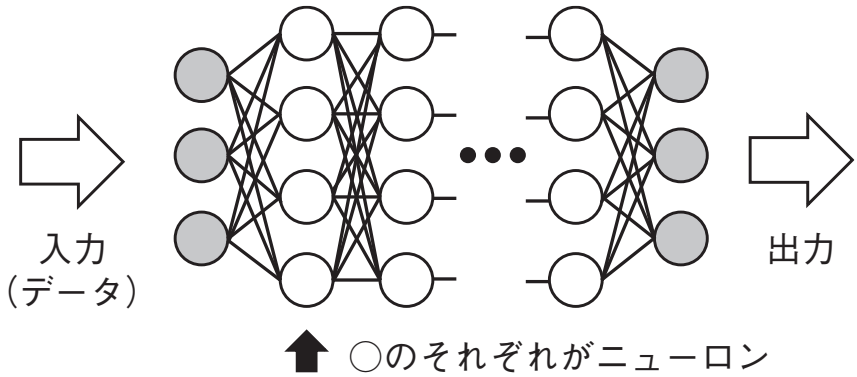

図２　何層も重ねられた人工ニューラルネットワーク

ある」という出力を返すという具合に，適切な分類ができるようになる．このような人工ニューラルネットワークの調整こそ，AI がデータにもとづいて学習をする，ということの正体である．

近年では，コンピュータの性能の向上により，人工ニューラルネットワークを何層にも重ねられるようになった．この「何層も重ねられている」ということこそ「深層学習」という語の「深層」が意味していることである．以前よりもはるかに深く層を重ねることができるようになったために，現在では AI の情報処理の効率が指数関数的に増大し，そのおかげで分類をはじめとする機能が高性能になったことから，AI の実用化が進んできたわけである[1)]．

(2) 科学的探究での実用の事例

次に，AI が科学において実際にどのような役割を果たしているのかを見ていくことにしよう．現在 AI は，科学的探究のさまざまな場面で発見や予測を導くツールとして導入されている．その大きな要因として，非常に大規模なデータ，いわゆる「ビッグデータ」が扱えるようになったことが挙げられる．医学関連の分野と基礎科学における事例を通じて，この点を確認してみよう．

まず，医学分野では，病気の検出や発症の予測に AI が役立てられる．たとえば癌に関する

1）深層学習の技術的な概観として，岡谷（2015）などを参照．ここで述べたニューラル人工ネットワークの層の深さに加えて，深層学習とりわけ DCNN（Deep Convolutional Neural Networks，畳み込み深層ニューラルネットワーク）が1980年代から1990年代の「浅い shallow」人工ニューラルネットを凌駕している理由として，以下のような点が挙げられる（cf. Buckner 2019）．①浅い人工ニューラルネットでは使われている関数が単一だが，深層学習では複数の関数を組み合わせていることで，特徴量（画像認識でいえば対象の輪郭など）の抽出能力が強化された．②ニューロンをすべて連結させるのではなく，層の間でスパース（まばら）に連結するという技法により，学習すべきパラメータの数を劇的に減らしたことで学習効率が向上した．③学習データにノイズを混ぜたりニューロン間の結合の一部をわざと不活性にしたりする手法を用いることで，いわゆる過適合（over-fitting つまり特定のデータセットに適合しすぎて一般性を欠いた状態）の軽減や防止が可能になった．

ビッグデータを集めることができれば，それにもとづく訓練を経て学習したAIが癌を検出して腫瘍の位置を特定したり，あるいは，今後しかじかの確率でどこに何年以内に癌が発症すると予測したりすることができる（cf. 山口 2021）．関連する領域として，薬学における創薬の分野でも，薬物の候補となる分子構造をAIに探索させる，といった活用例が挙げられる（e.g. 種石他 2017）．

次に，基礎科学においては，物理学や天文学などに顕著な事例が見られる（cf. Sejnowski 2018: ch. 1）．こうした分野でも，実験から得られた大規模なデータが取り扱えるようになったことから，AIが強力な研究ツールとして使用されるようになった．素粒子衝突実験では，スイスのジュネーヴなど町ひとつ分の大きさにも相当する非常に巨大な装置で実験を行い，そこからビッグデータが得られる．大型望遠鏡を用いて，夜空の隅々まで観測することにより巨大なデータが手に入る．この種のデータは規模が大きすぎて人間には処理しきれないけれども，AIを使えば活用できるようになる．大型望遠鏡が与えてくれるビッグデータ――重力レンズの効果のせいで地球まで歪んだ経路で届くような遠くの天体からの電磁波のデータも含まれている――の中から遠方の銀河を見つけ出すには，“干草の中にある針を探す”といわれるほど手間のかかる探索作業が求められるが，そうした作業でもAIなら高スピードかつ自動的に実行してくれる（Young 2017）．ほかには，AIで高速化した近似計算による三体問題へのアプローチ（Breen et al. 2019）なども，基礎科学における活用の一例である．

第2節　ブラックボックスの懸念

深層学習を用いたAIが科学的探究に大きく貢献する場面は今後さらに増えていくだろう．しかし，前節でみた事例も含めて，実はそこには人間には理解できないブラックボックス的な性格が存在している．AIによる発見や予測がなぜ成功するのかについては，AIを使用している科学者のみならず，AIの専門家でさえも捉えがたい部分が残ってしまうのである．AIの導入が，従来の科学的探究におけるコンピュータの利用，およびそれに伴う自動化と大きく違うのはこの点だといわれる（呉羽・久木田 2020: 153–4でもこのことは指摘されている）．

分光計のような既存の装置と比べてみよう．なるほど，分光計の使用者は必ずしもそのメカニズムを十分には理解していないかもしれない．しかし，科学者は全体として共同体を形成しており，その中には分光計の原理や働きを熟知した専門家が存在している．つまり，こうした装置については，知の分業体制が成立しているおかげで，たとえ個々の使用者がそのメカニズムを把握しきれていなくても，どこかで人間の十全な理解のもとに置かれているのである．ところが，以下で見ていくように，AIを使った科学的探究には，AIの専門家にもそのメカニズムがよくわからないまま研究が進んでしまいかねない．

この点に関しては，科学からいったん離れて，世の人々に衝撃をもって受け止められた「事件」を通じて確認してみたい――碁の世界的チャンピオンを打ち負かしたアルファ碁

（AlphaGo）の出現である（Silver 2018, cf. Sejnowski 2018: ch. 1）．当初，碁はきわめて複雑なので，チェスや将棋ではAIが人間の王者に勝てても，碁で勝利するのはしばらく先のことだと思われていた．だがアルファ碁は，韓国の世界トップ棋士のイ・セドルと対戦し，そうした事前の大方の予想を覆して勝利を収めたのである．深層学習による訓練を経たアルファ碁は，そのつどの盤面が自分や相手にとってどのぐらい有利・不利なのかを数値で評価し，次にどんな手を打てばその状況が改善するかを膨大な可能性の中から探索する，といったことが可能である．あるチャンピオンは，この対戦でアルファ碁が打った手について，それが非常に美しく，この手を人間が打つのは見たことがない，という趣旨のことを述べている．アルファ碁には，人間に打てるような手を上回る手を発見することができたのである．

アルファ碁とその後継バージョンの開発についても簡単に経緯を辿っておきたい．アルファ碁を開発したDeepMind社は，2016年にイ・セドルに勝利したあと，さらにバージョンアップしたAIをふたつ開発している．初代のアルファ碁は，人間がこれまでに打って蓄積されてきた16万もの棋譜（要するにビッグデータの一種）から学習をスタートしたが，学習がある程度進んだところで自分自身をコピーし，そのコピーと対戦を続けることでいっそう強さを増した．こうした訓練の結果として，アルファ碁はイ・セドルを降したのである．だが，その次に登場したアルファ碁ゼロ（AlphaGo Zero）は，棋譜のデータなしに碁のルールだけを教わり，そこから自分自身のコピーと対戦しながら学習を進めたところ，初代のアルファ碁に100対0の戦績で勝利するという強さを発揮した．さらにその次の2017年に開発されたアルファゼロ（Alpha Zero）になると，碁のルールさえも事前には教わらずに，自己対戦のみによる訓練だけで先代のアルファ碁ゼロに勝利する棋力をもつに至った．このように，一連のアルファ碁の後継AIの開発においては，最初に教えることが少なければ少ないほど強くなる，という逆説的な結果が得られた――DeepMind社の開発者たちは，これについて，もはや人間の知識の限界によっては制約されていない，といった趣旨のことを述べている（Hassabis and Silver 2017）．少し皮肉ともいえるが，人間が教えない方が強くなり，下手に人間が教えるとかえって弱くなってしまうのである．

AIのブラックボックス性が浮かび上がってくるのは，まさにこの点である．すでに述べたように，深層学習を用いたAIの学習とは，人工ニューラルネットワーク内のニューロン同士の結びつきの強弱を調整することを繰り返して訓練することだ．ところが，それがなぜアルファ碁をこれほど強くするのか，どうしてそんな美しく強力な手が繰り出されるようになるのかは，実はアルファ碁の開発者にさえも十分には理解できない．また，最初に人間が教えることが少ないほど強くなる理由もはっきりしない．あくまでもアルファ碁が結果として実際にチャンピオンに勝つほど――後継バージョンはそれ以上に――強いことが示されるだけなのである．これが，ブラックボックス的な性格がAIに見られるということの意味にほかならない．前節で見たパンダ画像の識別の例に即していえば，非常に複雑に結びついた多層の人工ニューラルネットワークが調整されていくことでAIの学習が進むのだけれども，その結果なぜうま

くパンダをパンダとして分類できるのかは人間にはわからない，つまりブラックボックスなのである．

このようなブラックボックス性が見られる理由をもう少し特定しておこう．（cf. Knight 2017, Schubbach 2019，瀧 2020，Durán and Jongsma 2021）．実はここで「ブラックボックス」と呼んでいるものは，ある意味では「ホワイトボックス」である．というのは，AI が行う情報処理の過程を逐一追うこと自体は原理的には不可能ではないからだ．しかしそれでも，人工ニューラルネットワークの構造の複雑さや学習に必要なデータの膨大さを考えれば直ちに明らかなように，情報処理の過程に現れる数字はあまりにも量が多いので，人間がその過程のすべてを見渡しながらそうした数字の意味や役割を理解することはできない．情報処理の細部だけに注目しても，ある特定の状態にある AI がその次にどのような状態に遷移するかさえ，関わる数字が多すぎて人間には予測することはできない．こうして結局のところ人間に得られるのはほとんど最終的な結果だけに限られてしまい，得られた結果がどれほど有用でも，その結果を導き出す過程やそれに先立つ学習プロセスにおいて何が起こっているのかは完全には捉えられない――このことを評してアルゴリズムが「不透明である（opaque）」とも表現される[2)]．以上が，深層学習を用いた AI にはブラックボックス的な側面が見られるということの意味である[3)]．

第 3 節　人間は科学から切り離されてしまうのか

AI の情報処理の過程が人間にとっては不透明であることは，碁に関してはとくに問題にはならないかもしれないが，科学的探究という場面では疑問が生じうる．AI で新たな発見なり精度の高い予測なりが行えるようになれば，確かに科学的探究は前進するだろうけれども，それがなぜ成功するのかの理由まで理解しなければ，科学という営みに不可欠の要素が決定的に失われてしまうのではないか．本節で見ていくように，こうした疑問は，人間と科学の関係に根本的な変更を迫る可能性についての懸念につながるものである．

(1) 予測と正しさの懸隔

ここでの問題を明確化するために，まずは科学史上の事例を取り上げてみたい――瀧雅人も

2）AI にまつわる不透明性は三種類に区分できると主張する論者もいる（Burrell 2016, cf. Carabantes 2020）．第一に，AI を管理している企業や政府などがユーザからその詳細を遠ざけているという意味での不透明性（法的規制等で対処可能），第二に，コンピュータのプログラムが読解できるリテラシーが社会に普及していないという意味での不透明性（教育によってある程度は対処可能），そして第三に，人間の認知能力とのミスマッチによる不透明性である．いうまでもなく三番目が本稿で論じているタイプの不透明性であり，容易な対処を許さない．

3）哲学者のボストロムはその著『スーパーインテリジェンス』で，人工ニューラルネットワークのこうした性質が，コンピュータの制御可能性にまつわる問題をいっそう悪化させる危惧について論じており（Bostrom 2014），これがいわゆるシンギュラリティ（技術的特異点）とその対処にも関連する問題に結びつくが，それについての検討は別の機会に譲りたい．

説明のために用いている紀元前2世紀半ばのプトレマイオスによる天文学の体系である．瀧も示唆するように（瀧 2020: 49），この事例から引き出せる教訓は，予測の精度が高いこと，すなわち，ある理論から精度の高い予測が導き出せるか否かということと，その理論がそもそも実際に正しいかどうかということは，別の問題だということだ．瀧の論点を敷衍・展開しながらこのことを確認していこう．

プトレマイオスの体系は，アリストテレスの学説を継承して洗練させた，いわゆる天動説（地球中心説）である．天動説を採るさいに問題になるのが，惑星の運動をどう扱うかという点だ．金星などの惑星を観測すると，恒星とは異なり，一定の方向に動く（順行）だけではなく，途中で止まったり（留），逆の方向に戻ったり（逆行）すること，つまり，文字通り「惑う」ような動きをすることがわかる．これは，現代の地動説から見ればごく当然の動きにすぎないが，静止した地球が宇宙の中心を占めているとする天動説ではその動きを説明するのは難しい．そこでプトレマイオスの体系では，地球を中心とする通常の円運動の軌道を設定したうえで，惑星はその円運動の軌道を中心とする小さな円の上を円運動する軌道を描くと解釈する．この追加された小さな円を「周転円」と呼び，そのおかげで惑星が止まったり，あるいは逆行したりすることも説明できるようになる（図3）．ただし，実際のプトレマイオスの体系は，さらに「離心円」「エカント」も加えたいっそう複雑な体系であった（Kuhn 1957）．

重要なのは次の点だ．近代科学が勃興し，ガリレオやコペルニクスといった人物の登場により地動説が確立されていくものの，惑星の運動に関する天動説の予測の精度は初期の地動説に引けをとらないものであった．つまり，天動説は正しくない理論であるにもかかわらず，一定以上の予測精度を実現していたのである[4)]．理論にもとづく予測や説明を観察可能な範囲の現

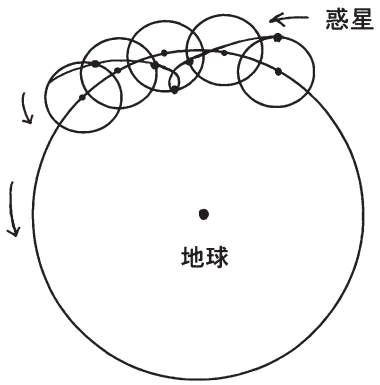

図3　プトレマイオスの体系（周転円のみ加えた簡略版）

4）これは科学哲学においてしばしば参照される事例である（Rosenberg 2005: ch. 4, ch. 6 などを参照）．ちなみに，プトレマイオスの体系の予測精度の高さは，惑星の運動が周期的であるがゆえに，円に円をいくつか重ねていけばフーリエ級数展開に近い仕方でその軌道をある程度まで近似できる，という理由から説明できる．

象とうまく合致させることは「現象を救う」と表現される（van Frassen 1980: ch. 3）．プトレマイオスの体系は，理論としては実際には正しくなかったけれども，まさしく目に見える現象――地上から見た惑星の運動――を救うには十分だったわけだ．

この例と AI を使った科学を比較するなら，ひょっとすると AI の働き方は，超高精度なプトレマイオス体系に似たものにすぎないのではないか，といった疑問が湧いてくる．AI は予測や検出などの能力がきわめて高いという意味で，なるほど非常によく現象を救ってはくれるのだけれども，理論として見ると世界を正しく捉えていない可能性がある．AI は，実際には存在しない周転円のような仕掛けを加えることで，現象を救う能力を高めているだけかもしれないのだ（cf. 瀧 2022: 49）．

他方で，両者には大きな違いもある．プトレマイオスの体系はやがてその誤りが判明したが，それに対し AI は，そのブラックボックス的側面のせいで，そもそも理論としての正誤を人間に見きわめることは不可能だ．予測に関していえば，正しいからこそ精度が高いのか，それともプトレマイオスの体系のように高精度ではあるものの誤っているのかを明らかすることができないのである．図 4 に示したように，法則と前提を合わせたもの（おおよそ理論と同一視できる）から予測や説明を導き出していくときに，AI を用いると図中の四角で囲んだ部分の中身が見えなくなってしまう．そこには周転円に似た要素が含まれているかもしれないのだが，ブラックボックス性により外部からのチェックが阻まれてしまうので，予測や説明などの最終的な結果だけが与えられるようになっているわけだ．

そしてこの点こそが，AI が従来にない決定的な変容を科学に生じさせる可能性と密接に関わっている．その可能性とは，呉羽真と久木田水生により提出され，また命名された論点である「異質な科学（alien science）」と科学からの人間の「疎外（alienation）」のふたつにほかならない（呉羽・久木田 2020）．以下ではこの論点に関する彼らの議論を確認していこう[5]．

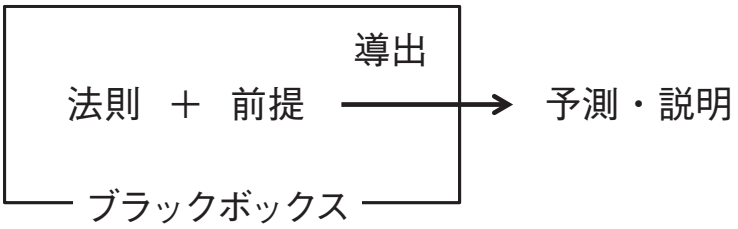

図 4　ブラックボックスからの予測と説明

5）一部だが以下では筆者の観点も交えている．

(2) 異質な科学

異質な科学，すなわち，人間がこれまで実践してきたような科学とは大きく異なる科学とはどのようなことか．呉羽・久木田は，人間の科学では単純性が重視されてきたのに対し，AIによる科学ではそれが成り立たなくなることをポイントとして挙げている（呉羽・久木田2020: 154-5）．比較のために，ここでも科学史を参照し，従来の人間の科学があくまでも「人間の頭向け」の科学でしかなかったことを確認しておきたい．

ニュートンによる万有引力の発見が偉大だったのは，リンゴの落下も月の運動も，すべてひとつの基本法則――万有引力の法則――にいくつかの前提を加えれば，それぞれについての予測と説明が導出できることを示した点にある．こうした古典的な科学理論に顕著なのが，そこに含まれる基本法則がごく少数であること，そしてそれゆえに「人間の頭に優しい」構造をしていることである．このことは，瀧（2020: 50）の表現にならえば「説明上の還元主義」が成立しているともいえる．なぜそう予測できるのか，なぜこれがこう起こったと説明できるのか，と問うてみよう．それは，かくかくの基本法則があり，そこにしかじかの初期条件を前提として加えれば，そうした予測や説明が導かれるからだ，という具合に，答えは法則と前提に帰着させられる．その意味で，予測や説明は少数の法則と前提に還元可能であり，有限な人間の知的能力でもそれなりに扱いやすいのである．

こうした構造は，「オッカムの剃刀」と呼ばれる科学の方法論的原則とも合致する[6)]．ここで先に述べておいた呉羽らが挙げる単純性というポイントが関わってくる．オッカムの剃刀とは，科学理論を作るときに不必要に多くの要素を増やしてはならない，という原則のことである．理論は単純であればあるほど，とりわけ基本法則と前提の数が少なければ少ないほど優れている――こうした単純性や倹約性の理論的な長所について述べたのがオッカムの剃刀という原則であるともいえる．その背景にあるのは，おそらくひとつには，単純な理論は人間の頭での理解のしやすさにも貢献するという点だと考えられる．法則や前提，そこに含まれるパラメータなどの要素の数が増えていくと，それに伴って理論の複雑さも増大するため，人間の知的能力を圧倒して容易な理解を許さなくなるのは目に見えている．

しかし，オッカムの剃刀は，AI を活用した科学的探究では必ずしも適当な方法論的原則ではないのかもしれない．AI を使えば，非常に大規模なデータをそのまま扱えるようになるので，説明上の還元主義に立たなくても，いわば複雑なものを複雑なまま捉えられる科学が成立する可能性があるからだ．この点についての示唆として，改訂を重ねている人工知能の代表的な教科書からの引用を記しておきたい．

> “深層ニューラルネットワークは，たとえそれがきわめて複雑であっても，しばしば非常にうまく一般化を行う．そうしたネットワークの中には，数百万ものパラメータをもつも

6）オッカムの剃刀に関する科学哲学の観点からの詳しい検討については，Sober（2015）などを参照．

のもあるほどだ”.（Russell and Norvig 2020: 655）

これまで単純性こそが優れた科学理論の証しだ――それゆえオッカムの剃刀に従え――と主張されてきたのに対して，AI を使った科学では，複雑さを縮減することなくそのまま受け入れられると見込まれるのである．そうなれば，単純性や倹約性と密接に関わる説明上の還元主義も，瀧（2020）にいわせれば，その反対の「ホーリズム（holism）」つまり「全体論」に取って代わられる可能性が出てくる．少数の基本法則に帰着させることなく全体を全体のまま捉えようとするホーリズムが――還元主義のようには人間の頭向きではないけれども――AI を使った科学の趨勢になるのかもしれない，というわけである．

だがそれがもたらすのは，人間がこれまで自分たちの頭向けに築き上げてきた科学理論とは大きく異なる世界認識のあり方なのではないだろうか．そうした新たな世界認識が人間にも共有できる保証はないし，まして AI のブラックボックス性を考えれば，そこでの理論がそもそもいかなる構造を備えているのかさえも窺い知ることができないかもしれない．こうして，呉羽と久木田が論じるように，AI の導入が人間の理解を超えた異質な科学を生み出してしまうのではないか，との懸念が生じるのである（呉羽・久木田 2020: 132–3，152–5）．

以上の懸念は，次の箇所で見る科学からの人間の疎外にも直結する．よく知られているように，マルクスは自分の労働の成果である生産物が労働者の手を離れて，むしろ人間性を圧迫するようになる「労働の疎外」を説いた．AI を用いた科学にも，それと似たような意味での疎外，すなわち，科学的探究を通じて作り出された知識が人間の手から離れてわれわれのものではなくなってしまう，という事態が生じるかもしれない．

(3) 科学からの人間の疎外

科学からの人間の疎外という可能性を考察するうえでは，そもそも科学と人間はどのような関係にあるのか，とりわけ人間にとって科学の意味や目的は何なのかを反省的に吟味してみる必要がある．科学の意味や目的をめぐる論争は必ずしも決着してはいないものの，呉羽らが検討している論点のうちでは，知的好奇心の充足，知的創造性の発揮，そして技術的な応用は，とりわけ重要な候補であると思われる（呉羽・久木田 2020: 123，152，156，162–3など）．ここでは科学的探究に AI を導入することでそれぞれに及ぶ影響について順に確認していこう．

まず，知的好奇心の充足，つまり世界のありようを科学が明らかにしてくれることでもたらされる興奮や喜びや満足は，AI の導入のせいで生じにくくなるかもしれない．それは何よりも，知的好奇心の充足には理解が大きな要因として必要だからである．だが，すでに上の（2）で触れたように，AI のもたらす世界認識が人間の手の届く範囲にあるとは限らない．

次に，知的創造性の発揮については，ある程度まで実現できると考えられる一方で，呉羽らはおおよそ以下のような疑問を提出している（呉羽・久木田 2020: 137–49）．知的創造性を新しく有用なアイデアや人工物を生み出していく能力という意味で捉えるなら（cf. Boden 2014），

先述の AI を活用した創薬のケースなどはそれに該当する．だがそうしたケースは，人間が知的創造性を発揮してなされた達成ではなく，あくまでも機械の自動的な過程の所産として捉えるべきではないか，という反論もごく自然で説得的である．知的創造性の本性をめぐる哲学上の議論を参照する場合でも，何らかの意味での達成という性格が認められない限り，そこに知的創造性を見てとることは難しい（とりわけ徳認識論における議論ではこの方向性は顕著である．e.g. Baehr 2018）．というわけで，AI の使用が科学的探究において知的創造性の発揮に寄与するという主張は，少なくともそのままでは，簡単には受け入れられてもらえないだろう．

最後に，技術的な応用について検討しよう．ここでの科学の技術的応用とは，ある対象のふるまいを科学理論にもとづいて予測することができるようになれば，その対象を人為的に制御することも可能になる，ということである．高精度の予測をもたらす AI がこの点に大きく寄与することは疑いないが，それでもブラックボックス性にまつわる問題に突き当たるのは避けがたい．呉羽らはこうした点について「人々が不安を抱くことが予想される」（呉羽・久木田 2020: 156）と述べている．われわれにとって不透明なアルゴリズムから導かれる結果を，果たして本当に技術的応用の場面に移してよいのか，という疑念は容易には払拭できない．

以上を科学からの人間の疎外という点からまとめよう．知的好奇心の充足および知的創造性の発揮の両者は，基礎研究に典型的に見られるように，たとえ実用面ではすぐには役には立たなくても，人間の知的・文化的な営みとしての意味を科学的探究に十分に与えていると考えられる．しかし，呉羽と久木田が論じているように，AI の導入により，将来的にそうした意味が科学から剥奪され，人間がこれまで科学との間に取り結んでいた関係が断たれてしまうのではないか，というのがここでの危惧である（呉羽・久木田 2020: 156）．また，AI がもたらす科学の技術的応用に関しては，予測と制御という目的はうまく達成されるのだから，ブラックボックス性のせいで理解はできなくとも，とにかく「鵜呑みにしなさい」「結果だけ信じなさい」という態度を人間に植え付けてしまいかねない．そのとき科学は，呉羽らが述べるように，「よく当たる『お告げ oracle』」（呉羽・久木田 2020: 155）を下すだけのものになってしまい，それは人間との隔絶にほかならないのではないか，といった疑問が湧いてきても不思議ではない．彼らも引用されているように（呉羽・久木田 2020: 156），こうした事態について“面白い可能性は「理解可能な」科学の時代をわれわれは閉じようとしているということだ”と述べる研究者もいるが（Bohannon 2017），それは同時に，科学からの人間の疎外という懸念と表裏一体だというべきだろう．

第 4 節　説明可能・理解可能な AI

前節で明確化した懸念のいくつかは，AI を用いた実際の科学的探究の進展とともに，それがどれほど現実的な可能性なのかが徐々に明らかになっていく性質のものである．その意味では，結局のところ杞憂に終わるかもしれないものだ．異質な科学にまつわる懸念のうち，理論

の構造があまりにもホーリスティックで人間には理解できないかもしれない，という可能性はまさにその一例である．だがそれでも，異質な科学の誕生であれ，科学からの人間の疎外であれ，AI のブラックボックス性が人間と科学の関係を何らかの仕方で変えてしまう見込み自体は十分ある．実際，近年では，AI のブラックボックス性に対応するために，表現は様々だが「理解可能（understandable もしくは intelligible）」「説明可能（explainable）」「解釈可能（interpretable）」な AI を作ろう，という研究の方向が現れ始めている（「説明可能な AI」は“XAI”とも記される）．本節ではそうした方向をふたつ取り上げて概観し，手短に検討を加えることにしよう．念のため注意しておけば，以下では考察の重心が，AI を導入した科学がもたらす懸念から，AI そのものについての科学，もしくは AI を含むシステムについての科学に移ることになる．

(1) 重要な意思決定における理由の提供

ひとつめの方向は，人間が最終的に重要な意思決定を行う必要がある状況で，AI がその助けとなるように使用される局面に関わっている[7)]．医療において治療方針の決定をどう行うかを考えねばならないときがそうした局面の例である．具体例として，MIT とマサチューセッツ総合病院が共同研究として取り組んでいる，AI を用いた乳癌の初期兆候の検出を取り上げよう．乳癌の場合，初期兆候が検出されたら，それを切除するか否かという重い意思決定を下す必要に迫られる．ところが，AI が教えてくれるのは「90％の確率で乳癌である」といった結果だけである．あるいは「60％の確率で今後20年以内に発症する」といった予測だけを結果として伝えられるかもしれない．いずれにせよ，そうしてただ単に診断の結果しか AI からは与えられず，その情報処理の過程がブラックボックスなのであれば，切除も含めた治療方針の決定を下す際，患者や医師に躊躇があっても仕方がないだろう（cf. Castelvecchi 2016）．そこで，最終的な結果を導き出すに至った推論の過程や診断の理由を説明してくれる AI の開発が目標として設定されるようになっている．

同様に重要な意思決定が求められる局面として，軍事作戦の遂行が挙げられる．ここでは，米国の DARPA（国防高等研究計画局）の研究事例を見ておきたい．DARPA は，兵器の開発に使える軍事研究を進めながら同時にその技術が民間でも利用できるような，いわゆるデュアルユース（軍民共用）の開発方針を採っている組織である．その最も有名な産物は，おそらくインターネットと GPS（全地球測位システム）だろう．こうした技術は，もともと軍事用に開発していたものが一般にも解放されて，現在では全世界で民生用に活用されるようになった

7）以下の事例についてはとくに Knight（2017）を参照．本文中で挙げた事例以外に，AI が重要な司法上の意思決定に関わる事例のひとつとして，米国で導入されている COMPASS（Correctional Offender Management Profiling for Alternative Sanctions）というシステムが挙げられる．COMPASS は再犯の見込みを予測することで裁判所の意思決定を支援するのだが，当然ながらそれに過度に頼る――人間の検察官や陪審員よりも重視する――ことに対しては危惧も表明されることになる（cf. Coekelbergh 2020: ch. 1）．

ものだ．いうまでもなく，デュアルユースを目指した技術開発という方針をめぐっては議論や賛否の声が絶えないが，ここで確認しておきたいポイントは，DARPA でも説明可能な AI 研究プログラムを進めているということである．

軍事で AI を利用するとき，以下のような由々しき状況が想定される．航空機や人工衛星で高空から撮った膨大な監視データをもとに「95％の確率でこの建物にはテロリストが潜んでいる」と AI が教えてくれる．それにより目標を定めて（しばしば自動操縦の）航空機からの攻撃を決めるわけだが，時折起こるように，民間人を誤って殺害してしまう事態も生じる．そう考えると，人間の兵士や分析官が軍事作戦の遂行についての最終的な意思決定を行うにあたっては，AI が提供してくれる結果だけに頼るということはしたくない．そこで，結果に至った過程や理由も何らかの仕方で AI が示してくれるようになる方が望ましい．このような事情から，DARPA でも，重要な意思決定を人間が下す必要がある局面で理由も説明してくれる AI の開発を目指した研究が始まっているのである．

とはいえ，こうした説明可能・理解可能な AI の研究の方向にも問題がないわけではない．AI が扱うデータはもともと非常に大規模であるため，実際に行われた情報処理の過程を人間に理解可能な形での理由の説明に変換する中で，過度の単純化が施されてしまっていても不思議ではない．だとすると，そこで重要な情報が欠落してしまい，そのせいで人間の意思決定に望ましくない影響が生じる，という可能性も見過ごせなくなる．この危惧は，AI の説明は本当の理由を提示しているのか，それを確かめる手立ては存在するのか，という疑念として表現することもできる．かりに説明可能・理解可能な AI の信頼性を調べる手段として何らかの別の AI を用いるなら，今度はその新たな AI の信頼性――その AI が本当の理由を示してくれているのか――をどう調べるのかという問題が持ち上がってくる．では，その解決のためにさらに別の AI を用いればよいのだろうか．いうまでもなく，ここに無限後退の臭いを嗅ぎつけるのはたやすい[8)]．

これと関連する点が，当の DARPA が刊行した文書ではこう述べられている．

> “機械学習の性能（予測の精度）と説明可能性の間には固有の緊張が存在している．たいていの場合，最も性能のよい手法（たとえば深層学習）が最も説明可能性が低く，そして最も説明可能性が高い手法（たとえば決定木）では，あまり精度が高くないのである”．（DARPA 2016: 7）

要するに，人間にとっての理解可能性と予測精度は，現状では技術的なトレードオフの関係にあるわけだ．この点を考慮すると，求められる意思決定が重要であるほど高精度の予測が必要になるとすれば，その分だけ人間が得られる説明可能性は低下してしまうことになる．このよ

8）こうした無限後退の可能性は Carabantes（2020）で手短に触れられている．

うに，ここで見た研究の方向は困難な課題を抱えており，少なくとも現状では発展の途上にあるというべきである．

(2) 重要な特徴の可視化

それではもうひとつの方向，すなわち重要な特徴の可視化を目指した研究に移ろう．ここでは再び図 1 でも示したパンダの画像認識を例に取り上げることにしたい（Goodfellow et al. 2014，cf. 瀧 2020）．画像データを入力されると，AI は「これは99.99％の確率でパンダだ」という結果を出力するが，なぜそう判断したのかは人間にはわからない．そこで，入力した画像のどこに注目してパンダと判断したのかを AI に視覚的に表現させるようにする．そのためのアルゴリズムを用いると「なるほど画像のこのあたりに注目してパンダと判断したなら納得できる」と，人間が目で見てそれなりに理解可能な画像が提示されるのである．この研究はある程度の成果を上げつつある．

もっとも，これにも課題がないわけではない．AI の画像認識の過程では，人間とは異なる仕方での視覚処理が生じうることもわかっている．元の画像に「敵対的ノイズ」と呼ばれる要素を混ぜたり色を少し変えたり，あるいは「敵対的サンプル」と呼ばれるデータを与えたりすると，途端に AI の判断の正答率が下がる場合がある．たとえば，元のパンダ画像に非常に薄いノイズを重ねると，人間の目には元の画像とまったく差がないように見えるにもかかわらず，AI はパンダではなく80％以上の確率で羊だと認識してしまう．さらに色味を調整すると，人間には相変わらずパンダに見える画像でも，AI は「51.0706％の確率でテディベアである」などと判断してしまう[9)]．

というわけで，AI が入力された画像のどこに注目しているのかを可視化することが，常に人間の理解に資するのかは，現在では不明である[10)]．さらに根本的な問題として，視覚ないしは知覚に限らず，そもそも深層学習は人間や動物の行う学習とは大きく異なるという重大な指摘がある[11)]．たとえば，AI は人間や動物が同様の情報を学ぶよりもはるかに多くの事例（データ）にもとづいて膨大な訓練を積まなくてはならない．また，深層学習で用いられる逆誤差伝播法（バック・プロパゲーション）という学習手法は，動物の脳の働きには見出しがたい．必要な訓練の量的相違については著名な神経科学者であるドゥアンヌ（Dehane 2020）によって

9）ただし，敵対的ノイズや敵対的サンプルの AI による処理の仕方が本当に人間の知覚と著しく異なっているのか，という疑問が近年の研究では提出されている（Elsayed et al. 2018）．これに関してバックナーは，AI は“人間の知覚的類似性のモデル化には成功しているかもしれないが，何かが**そのように見える**ところのもの（what something *looks like*）と何かが**実際にそうである**ところのもの（what it *really is*）との境界線を引くためのリソースをまだ手にしていないのかもしれない”（Buckner 2019: 14，強調は原文に従う）と興味深いコメントをしている．

10）この論点については Castelvecchi（2016）をも参照．

11）こうした論点とその手短な検討については Buckner（2019）を参照．なお，バックナーが他に挙げている哲学的・認識論的に興味深い論点として，概念の獲得メカニズムや，生得説と経験論の対立などがある．

も同様の指摘がなされており，多くのデータに触れねばならない AI とは対照的に，人間の乳幼児はある単語を学ぶのに 1 回か 2 回も触れれば十分身につけてしまう．ドゥアンヌは，人間をはじめ動物の学習に見られるこうした際立った効率のよさには，AI にはない認知的・生理的な機能，とりわけ注意や睡眠の働き，好奇心による動機づけ，そして人間の場合には白昼夢の存在が深く関わっている，と論じている．

以上を踏まえるなら，知覚や学習に関する AI の認知的な情報処理が，人間や動物について判明している――したがって今のところわれわれが理解している――範囲からは大きく隔たっているとしても不思議ではない．したがって，上で述べたような可視化の試みによって，人間とは異なる AI の視覚処理，ないしは学習を含む認知的な情報処理の一般的な本性にアプローチする試みは，ごく限られた成功しか見込めないかもしれない．

(3) 小　括

以上，AI のブラックボックス性にまつわる懸念への対応策として，ふたつの研究の方向を確認した．第一に，医療や軍事のように人間が重要な意思決定を迫られる場面で，理由や推論過程についての説明を提供することで判断の支援をしてくれる AI の開発を目指す研究の方向，そして第二に，画像認識に寄与する重要な特徴を可視化することで，人間にも AI の判断が理解可能になるようにしよう，という研究の方向である．

ここでは，このふたつの方向にさらなる明確化を施し，本節の内容をまとめておきたい．そのためには，ベルリン工科大学で機械学習を研究するモンタヴォンらによる「解釈／解釈可能性」と「説明／説明可能性」の定義が役立つと思われる（Montavon et al. 2017, cf. Carabantes 2020）．

モンタヴォンらの定義によると，まず，解釈とは，抽象的な概念を人間にもわかるドメインに写像することである．抽象的なベクトル空間や未知の文字・記号からなる列が解釈不可能なドメインにあるのに対し，画像（ピクセルの配列）やテキスト（語の列）が解釈可能なドメインにあるものの例だ．図 5 に示すように，ある対象のもつさまざまな特徴について膨大な数値

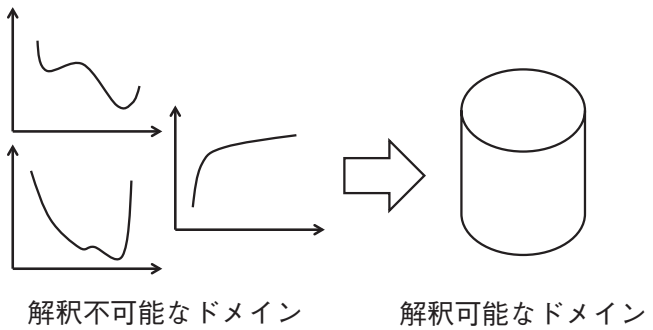

図 5　解釈不可能なドメインから解釈可能なドメインへの写像

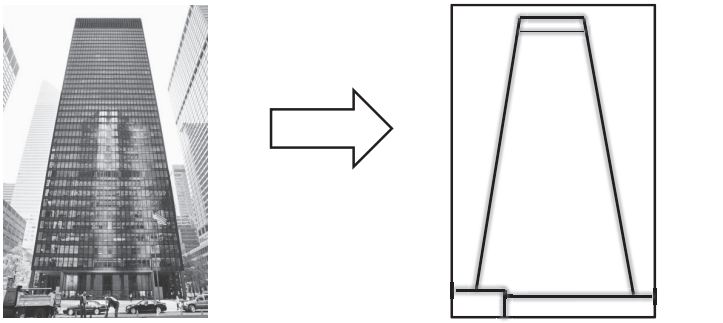

図 6　ヒートマップによる説明についての模式図
（図の左の写真は Wikimedia Commons からのものである：
https://commons.wikimedia.org/wiki/File:NewYorkSeagram_04.30.2008.JPG）

データを集めただけでは人間には解釈できないが，そのデータを画像に置き換えれば，元の対象がどんな形をしているか――ここでは円柱の形――は解釈可能になる．

次に，説明とは，解釈可能なドメインの中で，対象についての決定（たとえば分類や回帰）を生み出すのに寄与している特徴の集合である．2 次元データ（行列）の個々の値を色や濃淡として表現した可視化グラフの一種を「ヒートマップ」というが，入力画像のうち分類の決定に最も強く寄与したピクセルが強調されたヒートマップが，この意味での説明の典型例である．図 6 は，与えられた写真（画像）の中での人間とそれ以外の物体を分類するうえで，どの特徴がその決定に寄与したかを説明する様子を模式的に表している．

モンタヴォンらの定義に従えば，DARPA で行われている研究は，理由というテキストを AI に生み出させようとしている点で，解釈可能な AI の開発を目指している．これに対し，パンダ画像の分類における重要な特徴の可視化を目指した研究は，ヒートマップとおおむね同じ働きをする画像を作成させようとしている点で，説明可能な AI の開発を進めているのである[12]．

とはいえ，すでに述べたように，理解可能性と性能のトレードオフや，人間の脳との情報処理の仕方の違いなど，いずれの方向についても課題は少なくない．もしかすると，理解可能（解釈可能・説明可能）な AI の開発を目指す研究は，やがて原理的に乗り越えがたい技術的な限界に突き当たってしまうのかもしれない――そうした可能性さえも否定できない．その場合，ブラックボックス性の問題は解決されず，そのため AI がもたらしかねない科学と人間の関係の根本的な変容は，不可避の事態となるわけだ．そこで次なる課題として，それでもなお

12）モンタヴォンらとは用語法は異なるものの，彼らによる整理とも重なる詳細な議論をリプトンが示している（Lipton 2017）．それによると，解釈可能性は（1）「透明性（transparency）」と（2）「事後的な解釈可能性（post–hoc interpretability）」に大別される．ここでは詳細に立ち入ることはできないが，リプトンはさらに（1）を①シミュレート可能性，②分解可能性（decomposability），③アルゴリズム的透明性，（2）を①テキスト説明（text explanations），②視覚化，③局所的説明，④事例による説明，にそれぞれ下位分類している．

科学と人間に何らかの結びつきが残るとしたら，それはいかなる関係なのか，という問いが検討されねばならない．

第5節　人間の営みとしての科学

前節末尾の問いを考察していくには，科学という営為の実態に即して，何よりも科学的探究を人間が行う活動として捉える視点が有効であろう．そうした視点のもとで，AIと科学が今後どのような関係を取り結びうるのかを探り出すことが本節の目的である．そのうえで次の最終節では，AI導入が進んだ先の科学と人間の結びつきについての展望を示すことで，本稿全体の結論としたい．

(1) 人間が行う活動としての科学の姿

はじめに，現在人間が行っている科学的探究の大部分が分業体制で営まれていることが強調されねばならない．すでに第二次大戦中のマンハッタン計画にも見出されるように，いわゆるビッグサイエンスの著しい傾向として，知的分業の大規模化が現在に至るまで進んできた．結果として，知的分業に参加している研究者のうちの誰かひとりが，その全体像を完全に把握しているという状態は今ではほとんど成立しなくなっている．そのため，当該分野での権威者や特定の理論・実験の専門家を何らかの仕方で信頼し，あるいは依存することが欠かせないのである[13)]．

注意したいのは，実はここにもある種のブラックボックスがあるということだ．他の研究者の頭蓋骨に収められた脳の活動を外部から窺い知ることはできないし，かりに中を覗くことができたとしても，現在の脳神経科学の水準ではその働きを十分に理解することはできない．あまりにも当然のことではあるが，個々の研究者は分野の権威者や自分以外の専門家の脳内で進行する情報処理の過程をつぶさに把握したうえで，信頼したり依存したりしているわけではない．にもかかわらず，これまでの科学的探究は正当な知の営みであると認められてきた．こうした意味で，もとより科学的探究は，その実態においてブラックボックス的な性格を備えた活動だったのである（cf. Shubbach 2019）．

その極端な例が，2012年にノーベル物理学賞を授与されたヒッグス粒子の研究である．この研究は超大型粒子衝突加速器を用いて実施されるため，非常に多くの人間の参加が必要であった．実際，受賞の理由となった論文には，およそ3000人もの著者が名を連ねている．ここでは明らかに，チーム内の誰かが研究の全体像を漏れなく隅々まで把握していることなどない．け

13）これは社会認識論における重要なトピックであるが（e.g. 戸田山 2002: ch. 11），そこでの議論をさらに掘り下げるなら，専門家とは誰なのか，いつどの専門家に何について頼るべきなのか，といった問いが検討されねばならない．専門家をめぐるこうした論点を扱った社会認識論上の古典的な論文としてGoldman（2002）を参照．

れども，お互いに頭の中がブラックボックスである研究者・専門家が互いに信頼・依存し合うことによって知の分業体制を成立させることで，研究は成功を収めたのである．

実のところ，これと類似の側面は，科学的探究に限らず日常的な認識の場面にも見受けられる．日常においてわれわれは他者を，多少は誤りの余地は受け入れつつも，それなりに信頼してはじめてまともな生活を送ることができているからだ．直接会った人の脳でも，テレビ画面上のニュースキャスターの脳でも，そこでの情報処理の過程はブラックボックス的である．にもかかわらず，そうした他者の証言をある程度までは信頼することで，われわれは日常生活で必要となる知識を形成しているのだ．それどころか，われわれは自分自身についてさえもブラックボックス的な側面を抱えつつ，なおも信頼せざるをえない，というのが実情である．目で見たり，耳で聞いたりといった知覚や，何かを思い出す記憶の能力について，自分の脳を覗き込んで情報処理の過程を確認したうえで信頼する，などということはまったく行っていない．人間は，自分も他人もアルゴリズムに不透明な側面が存在することを承知で，信頼や依存にもとづく知の分業体制に参加し，そこで科学的探究や日常的な認識実践を営んでいるのである[14)]．

(2) AI と科学の今後

科学の実態に関する以上の視点のもとでは，科学と AI の関わりについて，さしあたり以下のような主張や方針を導き出すことができるだろう．

① **現状維持**　科学的探究を含め，およそ人間の認識実践には，現在すでにブラックボックスが溢れている．そこに，AI という新たなブラックボックス要因をさらに組み込んだところで，従来と大きく異なる抜本的な変化は生じるとは考えられない．それゆえ，AI を用いた科学にまつわる様々な問題について，そもそも心配は無用だ．

② **AI を組み込んだ効果的な分業体制の探究**　ブラックボックス的な側面を有する人間の集団からなる知の分業体制の中に，同じくブラックボックス的な性格をもつ AI をどのように組み込めば効果的な科学的探究が実現するのか，を明らかにするような研究を進めるべきである．人間と AI のブラックボックス性には相違があるが，両者を知の分業体制に組み込んだ場合に，その相違がいかなる影響をもたらすかについても解明を目指す．

③ **AI による脳の解明**　ブラックボックス的な側面をもつわれわれ自身を探究の対象とし，もっと理解を深めた方がよい．それには，人間をはじめとする動物の脳のまだ解明されていない領域や機能の実態に AI を用いて迫る研究を推進せねばならない．

それぞれ手短に評価・検討してみよう．まず，現状を是認する①は，カリフォルニア大学の

14）証言の信頼性は，認識論においてはイギリス経験論以来の大きな検討課題のひとつであり，とりわけ近年ではインターネット上のソーシャルメディアの発達により重要性を増している．現代における古典的な著作としては Coady（1992）を，より最近の著作としては Shieber（2015）を参照．

機械学習の研究者であるバルディが実際に表明している見解と重なる．バルディの言葉を引くなら，科学者を含めて人間は“脳をいつも使っていて，いつでも脳を信用している．そして脳がどう働いているかはわかっていない”（Castelvecchi 2016: 23）のである．結局のところ科学者は，頭蓋骨の中にいつもブラックボックスを入れているのだから，あらためてAIの不透明性を気にする必要もない，というわけだ．確かにこれは，研究の実践においては引き受けざるをえないプラグマティックな立場かもしれないが，全体としては単なる開き直りにも聞こえる．けれども，だからといってそこで①を切り捨ててしまうのではなく，ではなぜわれわれはブラックボックスのはずの人間の脳を信用しているのか，という問い返すこともできるだろう．これは，たとえば認知心理学上の新たな研究プログラムの出発点となりうる問いであり，もし実際にそうなれば，①はそれを可能にした貴重な見解として再定位されることになる．しかもそれは，同じくブラックボックス性を備えたAIへの信頼という問題に関わる点で，次の②にも資するはずだ．

②はなかなか興味深い主張といえるだろう．そこでは，知の分業体制のもとで進む科学的探究の中に，AIをどのように配置するべきかを明らかにすることを目指す，新たな研究プログラムの構想が提唱されている．とりわけ，人間の集団とAI――さらに実験機器などの他の人工物――が結節点となって織りなすネットワーク的なシステムの有効性や最適化についての探究は，それ自体がAIの効果的な活用が期待できる研究領域として成立すると考えられる[15]．

最後の③は，脳の研究とAIの研究とを歩調を合わせて進めようという主張である．脳の研究でひとつ大きな問題になるのが，得られるデータが巨大になることだ．脳には無数のニューロンが存在し，その活動についてのデータを集めていくと，しばしば人間には手に負えない規模のビッグデータになってしまう．だが，第1節で見たように，むしろそうした場面でこそAIは活躍してくれる．また，統合失調症や認知症といった人間の脳の疾患についてのデータも集まりつつあるが，同様にそれが扱いづらい規模の大きさになっても，AIを活用することで対処が図れるだろう（Sejnowski 2018: ch. 12）．こうした研究は，従来はブラックボックス的だった脳の領域や機能に光を当て，その理解を前進させてくれると考えられる．さらに，そこで得られた知見の一部は，もしかするとAIのブラックボックス性についての研究にも，限定された範囲ではあっても応用できる可能性がある[16]．そうしてAIの研究が進展したら，今度はその成果を脳の研究に還流させればよい．このように，脳の研究とAI研究が互いを相補的に促進させていくような，新たな科学的探究を構想することができる――むろんこれはまだ思弁の段階にすぎないが，ひとつの可能性としては十分ありうる姿ではないだろうか．

15）ここでの議論はラトゥールらのアクターネットワーク理論を思い起こさせるが，立ち入った検討は別の機会に譲ることにしたい．

16）もちろん既存の神経科学的知見をAIに応用する研究もすでに存在する．一例として，動物が一般的に備えている注意や強化学習のような機能を担う構成要素をAIに組み込む方向での研究が挙げられる（Hassabis et al. 2017）．

第6節　結　論

これまでの考察を簡単にまとめたうえで暫定的な結論を示し，本稿を締めくくろう．冒頭で示したように，AIが科学的探究で担う役割は大きくなりつつある．ところが，一方でそのことは，ブラックボックス化の問題を核として，人間と科学の間に断絶がもたらされる懸念を生む．そうした懸念に対する対応策として，説明可能・理解可能なAIの開発を目指す研究も始まっており，ある程度の成果も上がっているものの，その成功は必ずしも保証されていない．しかしそれでも，第5節で考察したように，たとえば脳の研究とAIの研究とが相補的に進められるような，新たな領域を切り開いていける可能性は十分に残されている．

以上から，AIと科学と人間の三者関係の将来，とりわけAIが人間を科学から切り離してしまうのではないか，という懸念にまつわる問いには，次のような暫定的な解答が与えられるようになる．それは直ちに，全面的に起こるような事態ではない．むしろこれから当面のあいだ続くのは，AIを媒介にした新しい研究領域が次々と登場し，まさにそのことによってAIと科学と人間がいっそう緊密な関係を取り結んでいく将来である，と．

謝　辞

本研究は，関西大学経済・政治研究所の研究費支援を受けたものである．また本稿は，JST・RISTEX研究開発プロジェクト「人と情報テクノロジーの共生のための人工知能哲学2.0の構築」，および科学研究費補助金・挑戦的研究（萌芽）「コンピュータ化によるパラダイム変化とその対応」（課題番号：21K18351）にもとづく研究成果の一部である．

文　献

Baehr, J., Intellectual creativity. In B. Gaut and M. Kieran eds., *Creativity and Philosophy*, 2018.

Boden, M. A., *The Creative Mind: Myths and Mechanisms*, 2nd edition, 2004.

Bohannon, J., The cyberscientist, *Science*, Vol. 357, Issue 6346, 2017.

Bostorom, N., *Superintelligence*, 2014. 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』，日本経済新聞社，2017年

Breen, P. G. et al., Newton vs the machine: solving the chaotic three-body problem using deep neural networks, *ArXiv*: 1910.07291, 2019.

Buckner, C., Deep learning: a philosophical introduction, *Philosophy Compass*, 14: e12625, 2019.（https://doi.org/10.111/phc3.12625）

Burrell, J., How machine 'thinks': understanding opacity in machine learning algorithms, *Big Data and Society*, 2016.（https://doi.org/10.1177//2053951715622512）

Carabantes, M., Black-box artificial intelligence: an epistemological and critical analysis, *AI and Society*, 35, 2020.

Castelvecchi, D., The black box of AI, *Nature*, 538, 2016.

Coady, C. A. J., *Testimony: A Philosophical Study*, 1992.

Coeckelbergh, M., *AI Ethics*, 2020. 直江清隆他訳『AIの倫理』，丸善出版，2020年
DARPA（Defense Advanced Research Projects Agency），Explainable artificial intelligence（XAI），*DARPA-BAA*, 16-53, 2016.
Dehane, S., *How We Learn : Why Brains Learn Better Than Any Machine ...for Now*, 2020. 松浦俊輔訳『脳はこうして学ぶ――学習の神経科学と教育の未来』，森北出版，2021年
Durán, J. M. and Jongsma, K. R., Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI, *Journal of Medical Ethics*, 47, 2021.
Elsayed, G. et al., Adversarial examples that fool both computer vision and time-limited humans, *Advances in Neural Information Processing Systems*, 31, 2018.
Goldman, A., Experts : which ones should you trust? In his *Pathway to knowledge : Private and Public*, 2002.
Goodfellow, I., Shlens, J., and Szegedy, C., Explaining and harnessing adversarial examples, *ArXiv Preprint ArXiv* : 1412.6572, 2014.
Hassabis, D. and Silver, D., AlphaGo Zero : leaning from scratch. DeepMind com. 2017.
（https://deepmind.com/blog/alphago-zero-learning-scratch）
Hassabis, D. et al., Neuroscience -inspired artificial intelligence, *Neuron*, 95(2), 2017.
Knight, W., The dark secret at the heart of AI. *MIT Technology Review*, 2017.
（https://www.technologyreview.com/s/604087/the-dark-secret-at-theheart-of-AI/）
Kuhn, T. S., *The Copernican Revolution*, 1957. 常石敬一訳『コペルニクス革命』，講談社，1989年
Lipton, Z. C., The mythos of model interpretability, *Arxiv* : 1606.03490v3, 2017.
（https://arxiv.org/abs/1606.03490v3）
Montavon, G., Samek, W., and Müller, K. R., Methods for interpreting and understanding deep neural networks, *Digital Signal Process*, 73, 2017.
Rosenberg, A., *Philosophy of Science: A Contemporary Introduction*, 2nd edition, 2005. 東克明・森元良太・渡部鉄兵訳『科学哲学――なぜ科学が哲学の問になるのか』，春秋社，2011年
Russell, S. J. and Norvig, P. eds., *Artificial Intelligence : A Modern Approach,* 4th edition, 2020.
Schubbach, A., Judging machines: philosophical aspects of deep learning. *Synthese*, 2019.
（https://doi.org/10.1007/s11229-019-02167-z）
Sejnowski, T. J., *The Deep Learning Revolution*, 2018. 銅谷賢治監訳『ディープラーニング革命』，ニュートンプレス，2019年
Shieber, J., *Testimony : A Philosophical Introduction*, 2015.
Silver, D. et al., A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, *Science*, 362（6419）, 2018.
Sober, E., *Ockham's Razors : A User's Manual*, 2015. 森元良太訳『オッカムのかみそり――再節約性と統計学の哲学』，勁草書房，2021年
van Frassen, B. C., *The Scientific Image*, 1980. 丹治信春訳『科学的世界像』，紀伊國屋書店，1986年
Young, M., Machine learning astronomy, *Sky and Telescope*, December, 2017.
岡谷貴之『深層学習』，講談社，2015年
呉羽真・久木田水生「AIと科学的探究」，稲葉振一郎他編『人工知能と人間・社会』，勁草書房，2020年，所収
瀧雅人「騙されるAI」『日経サイエンス』2020年1月号
種石慶ほか「創薬とAIの良好な関係」，夏目徹編『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――研究に生産性と創造性をもたらすテクノロジー』，羊土社，2017年，所収
戸田山和久『知識の哲学』，産業図書，2002年
山口聡一郎「コンピュータによる医療画像診断の発展」，『セミナー年報2020』，関西大学経済・政治研究所，2021年