はじめに

人工知能：1958 年、1972 年、そして現在

初期の人工知能研究における中心的人物であったハーバート・サイモンとアラン・ニューウェルは、1958 年に、10 年以内に以下のようなことが実現するだろうという予測を立てた（Simon and Newell, 1958, pp. 7–8）。

・デジタルコンピュータがチェスの世界王者になる。
・デジタルコンピュータが数学の重要な定理を発見し、証明する。
・デジタルコンピュータが、高い美的価値をもつと批評家が認めるような音楽を作曲する。
・心理学理論の大半は、コンピュータプログラムという形をとるようになる。

しかし、その 10 年後にあたる 1968 年の時点では、いずれの予測も実現にはほど遠い状態にあった。人工知能研究のこのような状況について、アメリカの哲学者ヒューバート・ドレイファスは、1972 年に出版された『コンピュータには何ができないか』の初版において、つぎのように述べている。

> 伝統哲学の先入見を去って記述的あるいは現象学的な証拠に頼れば、どんな形の知的振舞いにもプログラム不可能な人間の能力が含まれている、という示唆が得られる。……人工知能が可能であるか否かという問いが経験的問いであるかぎり、認知のシミュレーションまたは人工知能において、これ以上の重要な前進はほとんど見込まれないという答えになると思われる。（Dreyfus, 1992, p. 285; 邦訳 pp. 487–488）

ドレイファスは、人工知能研究に対して一貫して否定的な立場をとり続けていたことで知られている。彼は、1970年代前半の段階で、人工知能研究にはこれ以上の重要な進展は見られないだろうという評価を下していたのである。

ドレイファスの評価に反して、人工知能研究は、その後も紆余曲折を経ながら発展を続けてきた。上に挙げたドレイファスの文章は、いわゆる第1次人工知能ブームが行き詰まりを迎えた時期に書かれたものである。その後、1970年代後半から80年代にかけて、エキスパートシステムなどの研究が活発に進められ、第2次人工知能ブームが到来した。1990年代になるとそのような研究も停滞期を迎えるが、その後の第3次人工知能ブームにつながる機械学習とニューラルネットワークの研究は、着々と進められた。1997年には、IBMのDeep Blueが当時のチェスの世界チャンピオン、ガルリ・カスパロフに勝利した。サイモンとニューウェルの予測は、30年遅れで現実のものとなったのである。2012年には、画像認識に関するコンテストで、深層学習を用いた画像認識システムが、従来の手法を圧倒的に上回る成績を挙げて注目を集めた。第3次人工知能ブームの到来である。その後の人工知能研究の急速な進展は、われわれが日々目にする通りである。Google社のAlphaGoは、2016年に囲碁の世界トップ棋士だったイ・セドルに勝利した。2023年には、ChatGPTをはじめとする大規模言語モデルが日々のニュースを賑わせている。このような現状をふまえれば、サイモンとニューウェルの見立てはやや楽観的すぎたにせよ、それほど的外れなものではなかったようにも思われる。

今日では、人工知能研究の見通しに関して、ドレイファスとは対照的な評価を目にすることも珍しくない。たとえば、東京大学の人工知能研究者である松尾豊は、一般向けの著書でつぎのように述べている。

> ディープラーニングは特徴表現学習の一種であり、その意義の評価については、専門家の間でも大きく2つの意見に分かれている。1つは、機械学習の発明のひとつにすぎず、一時的な流行にとどまる可能性が高いという立場である。これは機械学習の専門家に多い考え方だ。もう1つは、特徴表現を獲得できることは、本質的な人工知能の限界を突破している可能性があるとす

る立場である。こちらは機械学習よりも、もう少し広い範囲を扱う人工知能の専門家に多いとらえ方である。本書は、後者の立場に立つ。（松尾, 2015, p. 180）

別の本で、松尾はつぎのようにも述べている。

私は、知能には鳥が飛ぶことと同じように原理があり、それを工学的に利用することもできるはずだと思っています。すでにディープラーニングで最大の難所が突破されたいま、あとは身体性や記号操作の仕組みを獲得できれば、知能の原理の大方は説明がつくのではないか――それが私の考えです。（松尾, 2019, pp. 152–153）

さらにその先を見据えている人々もいる。たとえば、物理学者スティーヴン・ホーキングやテスラ社のイーロン・マスクは、近い将来、人間を上回る知能をもつ人工超知能（artificial superintelligence）が誕生し、人類は制御不能となった人工超知能によって滅ぼされる可能性があると考え、人工超知能の開発やその実現につながる研究には制約が必要だと主張している[1)]。

過去 50 年ほどのあいだに、なぜ人工知能に対する評価はこれほどまでに変化したのだろうか。もちろん、その理由は、この間に人工知能研究が飛躍的な進展を遂げたということである。では、なぜそのような飛躍的な進展が生じたのだろうか。一つの答えは、深層学習がその鍵だ、というものだろう。では、深層学習が登場する以前の人工知能とそれ以後の人工知能には、どのような違いがあるのだろうか。従来の人工知能研究が直面した困難は、深層学習によってすべて克服されたのだろうか。松尾が言うように、人工知能研究は最大の難所をすでに突破したのであり、人間のような知能をもつコンピュータの登場は、時間の問題なのだろうか。

1）たとえば以下の報道を参照。
https://www.bbc.com/news/technology-30290540
https://www.nikkei.com/article/DGXMZO41827570X20C19A2000000/（2024 年 1 月 7 日確認）

人工知能の哲学 2.0

知能は、人間の重要な特徴の一つである。それゆえ、デジタルコンピュータのような機械は知能をもつことができるかという問いは、知能とは何か、人間とはどのような存在かといった問いとも密接に関連する。このような理由から、人工知能研究の初期から、哲学者は人工知能研究に強い関心を抱いてきた。そして、人工知能研究には原理的な限界があり、コンピュータが人間のような知能をもつことは原理的に不可能であると哲学者が主張することも、珍しくなかった。その結果、人工知能の可能性と限界をめぐって、人工知能研究者と哲学者のあいだでは、活発な論争が交わされてきた。ところが、1990 年代に入り人工知能研究そのものが停滞期を迎えると、このような論争は次第に下火になってしまった。

21 世紀に入り、人工知能研究はふたたび爆発的な進展を見せている。では、哲学者がこれまで展開してきた批判は、すでに乗り越えられたのだろうか。汎用人工知能や人工超知能の誕生は時間の問題だという松尾やホーキングの見立ては、正しいのだろうか。これらの問いに答えるためには、人工知能研究の現状をふまえて、かつて行われていた人工知能をめぐる哲学的考察をアップデートする必要がある。具体的には、以下のような問いを検討する必要があるだろう。

- 従来の人工知能研究はどのような基本的発想に基づいているのか。
- 従来の人工知能にはどのような原理的問題があったのか。
- 従来の人工知能と現在の人工知能には、どのような違いがあるのか。
- 現在の人工知能は、従来の人工知能の原理的問題を克服したのか。
- 現在の人工知能にも課題や限界があるとすれば、それは何か。
- 現在の人工知能は、人間の知能を理解する上でどのような手がかりを与えてくれるのか。

本書は、これらの問題の検討を通じて、人工知能の哲学をバージョン 2.0 にアップデートしようという試みである。人工知能には何ができて何ができないのかという問いを主たる問いとする点で、人工知能の哲学 2.0 は、従来の人工

知能の哲学（「人工知能の哲学 1.0」）と共通の問題意識に立脚している。しかし、この問題を論じる文脈には、両者のあいだで大きな違いがある。1972 年にドレイファスが『コンピュータには何ができないか』の初版を出版したとき、人工知能研究は、当初期待されたような成果を挙げることができていなかった。それゆえ、ドレイファスの問題意識は、なぜ人工知能研究は期待された成果を挙げることができないのか、そこにはたんなる技術的課題を超える原理的な困難があるのではないか、ということにあった。人工知能研究の現状は、これとは対照的である。過去 10 年ほどのあいだに、人工知能研究は予想を上回る勢いで進展を見せている。人工知能にできないことはない、汎用人工知能や人工超知能の実現も時間の問題だと考える人も少なくない。本当にそうなのだろうか、ということをあらためて検討してみようというのが、本書の問題意識である。

*

本書は 4 つのパートからなる。第Ⅰ部では、古典的な人工知能の基本的発想とその原理的な問題を検討する。第Ⅰ部は、いわば人工知能の哲学 1.0 のおさらいである。第Ⅱ部と第Ⅲ部では、深層学習を中心に、現在の人工知能研究の基本的な手法を確認した上で、古典的な人工知能と現在の人工知能の違いや、現在の人工知能の課題と限界について考察する。第Ⅳ部では、現在の人工知能の課題と限界についてあらためて考察するとともに、人間の心や知能を考える上で、人工知能はどのような手がかりを与えてくれるのかを考察する。

本論に入る前に、注意点をいくつか述べておこう。

・人工知能について論じる際には、当然のことながら、人工知能とは何か、知能とは何かといったことが問題になる。この点に関して、本書ではさしあたり、人工知能研究を、知能をもつ人工物を作る試みと理解し、知能を、環境に対して適切な行動を生み出す能力と理解しておくことにする。そして、とくに区別が必要な場合を除いて、知能、知性、思考、認知といった語を、上のような意味での知能を表す語として区別なく用いることにする。
・本書では、artificial intelligence を基本的に「人工知能」と表記する。「ゲ

ーム AI」のように表現が定着している場合には「AI」という表記を用いることもあるが、意味上の区別はない。

・本書は人工知能そのものの教科書ではないが、必要に応じて、人工知能のさまざまな手法をある程度くわしく紹介している。それは、「私が考えるところの人工知能」ではなく、現実に存在する人工知能について意味のある考察を行うためには、考察の対象そのものについて学ぶことが不可欠だからである。（物理学を一切学ばずに物理学の哲学ができるだろうか？）

・特定の話題についてややくわしく論じている箇所や、やや脇道的な話題に触れている箇所には、見出しに＊をつけた。これらは読み飛ばしても本文の理解に支障はないが、本文からそのまま読み続けられるように配置してある。

・本書は 12 章構成だが、各章の分量にはばらつきがあるので、授業などで利用する際にはこの点に配慮する必要があるだろう。たとえば、哲学的な検討をおもに行っている第 3 章と第 10 章は複数回に分けたり、現在の人工知能の基本的な手法を紹介している第Ⅱ部を扱う際には別のテキストを用いて説明を補足するなどするとよいだろう。

読書案内

第 2 次人工知能ブーム期までは、人工知能の哲学に関するさまざまな論文や書籍が出版されていた。人工知能の哲学の概説書としては、つぎの 2 冊が代表的である。本書、とくに第Ⅰ部を執筆する上でも、これらの本はおおいに参考にしている。

・Haugeland, J.（1985）. *Artificial intelligence: The very idea.* MIT Press.

・Copeland, J.（1993）. *Artificial intelligence: A philosophical introduction.* Blackwell.

また、この時期の人工知能の哲学に関する代表的な論文を収録した論文集としては、つぎのものがある。