はじめに

鈴木貴之

第3次人工知能ブームが到来してから、10年ほどが経とうとしている。人工知能は、すでにわれわれの生活のさまざまな場面に浸透している。急速に発展する人工知能をめぐっては、さまざまな社会的論争も生じている。そこでは、人工知能によってわれわれの職の多くが失われてしまう可能性や、人間よりも高い知能をもつ人工超知能が人間に対して敵対的な行動をとる可能性のように、一昔前であればSF的と思われていた問題も、現実味のある問題として議論されている。とはいえ、われわれの身の回りを見渡してみても、SF映画に登場するような人工エージェントはまだ存在しない。人間のような知能をもち、自律的に行動する人工エージェントの実現には、まだしばらく時間がかかりそうである。

われわれは、この現状をどう評価すべきだろうか。われわれの社会や生をよりよいものにするためには、われわれは人工知能とどうつきあっていけばよいのだろうか。本書で考察したいのは、このような問いである。

人工知能研究の歴史

本書の問題設定を明確にする前に、まずは人工知能研究の歴史を簡単に振り返っておこう[1)]。

人工知能研究が本格的に始まったのは、1956年に米国で開催されたダートマス会議からだと言われる。このころ、当初計算機械として作られたデジタルコンピュータを用いて、思考する機械を作ろうという構想が生まれつつあった。

1) 人工知能研究の現状の概観としては、たとえば松尾（2015）；谷口（2020）；ミッチェル（2021）などを参照。

ダートマス会議では、このような試みに人工知能という名称が与えられ、以後、これがこの研究分野の一般的な名称となった。第 1 次人工知能ブームの始まりである。

初期の人工知能研究は、一般的な思考能力をデジタルコンピュータで実現しようというものだった。その代表的な研究は、ハーバート・サイモンとアラン・ニューウェルによる一般問題解決器（General Problem Solver: GPS）である。GPS は、演繹的推論を行ったり、さまざまなパズルを解いたりすることができたため、注目を集めた。しかし、GPS がうまく扱うことができるのは、ヒューリスティックを用いた探索という形にうまく形式化できる問題に限られ、また、そのような問題においても、人間が問題を適切な仕方で形式化する作業が不可欠であることが次第に明らかになっていった。文字通りの一般問題解決能力を人工知能に与えることは、簡単なことではなかったのである。

GPS の挫折をふまえて、その後の人工知能研究では、対象領域を限定した人工知能の開発が進められることになった。代表的な研究事例の 1 つは、テリー・ウィノグラードによる SHRDLU である。SHRDLU は、現実世界を単純化した小規模な世界であるマイクロワールドで行動する。マイクロワールドは色のついた積み木から成り立ち、SHRDLU は、命令に従って積み木を操作したり、積み木に関する質問に答えたりする。（マイクロワールドは仮想的な世界なので、実際に行うのは仮想的な操作である。）この時期の人工知能研究に見られたもう 1 つの方向性は、エキスパートシステムの開発である。これは、特定の問題領域に関して人間の専門家がもつ知識を人工知能上に再現することを目指したものである。代表事例の 1 つ MYCIN は、人間の医師が用いていると考えられるさまざまな知識を人工知能上に再現することで、血液感染症が疑われる患者の検査データから、患者がかかっている感染症と有効な薬品を特定するというものである。

これらの研究は大きな注目を集め、1970 年代から 80 年代にかけて、第 2 次人工知能ブームが起こった。しかし、このようなアプローチも次第に問題に直面するようになった。たとえば、SHRDLU のようなアプローチによって現実世界で行動するエージェントを作ろうとすると、世界に関する無数の常識を知識として身につけさせる必要があるということが明らかになった。小規模で限

定された問題領域において成功を収めたアプローチを、はるかに複雑な現実世界における問題解決に適用することは、困難だったのである。

これらのアプローチが行き詰まりを見せた結果、人工知能研究は冬の時代を迎えることになった。しかし、その間に、その後の第3次人工知能ブームにつながる研究も進展していた。第一の要素は、機械学習である。古典的な人工知能研究では、人間のプログラマーが、問題解決に必要な知識を人工知能に与える必要があった。しかし、必要な知識をすべて明示化することは、きわめて困難だった。これに対して、コンピュータにデータから学習させるのが機械学習である。1990年代以降、数理モデルに機械学習を適用するさまざまな手法が開発され、機械学習の有用性が広く認識されていった。

第二の要素は、ニューラルネットワークである。ニューラルネットワークは、生物の脳にヒントを得たアーキテクチャで、入力層の活性化パターンを出力層の活性化パターンに変換することを基本的な仕組みとする。ニューラルネットワークを用いた人工知能研究は1980年代に本格化し、画像認識や音声合成など、古典的な人工知能の手法ではよいパフォーマンスを示すことが困難だった課題において成功を収めたために、注目を集めた。とはいえ、1980年代のコンピュータ上でシミュレートできるニューラルネットワークは比較的小規模なものだったため、可能なことはかぎられていた。

2010年ころから、これら2つの要素が一体となり、深層ニューラルネットワークを用いた人工知能研究が一般的となった。そこには3つの背景要因がある。第一の要因は、コンピュータの性能向上によって、大規模なニューラルネットワークのシミュレーションが可能になったことである。第二の要因は、インターネットの普及によって、大規模な訓練データを用いた機械学習が可能になったことである。第三の要因は、大規模なニューラルネットワークで効果的に学習を進めるためのさまざまな手法の開発が進んだことである。これらの要因が重なったことで、深層ニューラルネットワークを用いた機械学習、すなわち深層学習を用いた人工知能が急速に進展し、第3次人工知能ブームが到来した[2)]。

2)　深層ニューラルネットワークの仕組みについては、第3章でより具体的に紹介されている。

深層学習が注目を集めた大きなきっかけは、2012 年に、畳み込みニューラルネットワークを用いたシステムが画像認識のコンテストで圧倒的なパフォーマンスを示したことである。2017 年には、深層強化学習を用いた AlphaGo が当時の囲碁の世界トップ棋士を破った。その後の人工知能研究の進展は、われわれが日々の生活で目にする通りである。

人工知能研究の現状と社会的・倫理的問題

深層学習の登場によって、人工知能研究は当初の目標を達成したのだろうか。現時点では、まだそう言い切ることはできないだろう。

第一に、現在ある高性能な人工知能は、いずれも特定の課題に特化したものであり、生物のように、知覚し、意思決定し、行動するといった働きをすべて兼ね備えるものではない。生物のように多様な課題を行うことができる知能は、しばしば汎用知能と呼ばれる[3)]。汎用性のある知能をもつ自律的なエージェントを人工的に作り出すことが人工知能研究の究極目標だとすれば、それはまだ達成されていないのである。また、深層学習によって汎用人工知能が実現できるかどうかも、いまのところ明らかではない。人工知能研究者の中には、汎用人工知能の実現のためには、深層ニューラルネットワークと記号的な古典的人工知能を統合することが必要だと考える人も多い。しかし、両者をどのように統合すればよいのかは、いまのところ明らかではない。

第二に、深層学習に関しては、理論的に解明されていない点も多い。たとえば、1 層のユニット数が十分に多ければ、2 層のネットワーク（入力層、1 つの中間層、出力層からなるネットワーク）で任意の関数を任意の精度で近似できるということが以前から知られている。ネットワークを深層にすることでなぜ性能が向上するのかは、明らかなことではないのである。また、数理モデルに関する従来の考え方によれば、モデルが複雑になるほど過学習が生じやすくなり、汎化性能（新しい事例に対して正しい出力を与える能力）は低下すると考えられ

3)　生物の知能は、特定の課題に特化したものではなく、知覚・意思決定・運動制御といった多様な課題に対処できるという意味では、汎用の知能である。しかし、人間以外の多くの生物の知能は、計算をしたり言語を用いたりすることができないという意味では、文字通りには万能ではない。その意味では、通常汎用知能と呼ばれるものは自律型知能と呼ぶのが適切かもしれないが、一般的な用法にならって、本書でもこのような知能を汎用知能と呼ぶことにする。

る。しかし、深層ニューラルネットワークはきわめて複雑なモデルであるにも関わらず、高い汎化性能をもつ。これは、従来の考え方からすればきわめて逆説的な事態である。このように、深層学習の原理に関しては未知の点も多い。これが意味することは、深層学習にはどのような注意点や限界があるのかもまた、まだ十分には明らかになっていないということである[4)]。

これらの理論的問題と並んで、現在の人工知能は、その性能が大きく向上したがゆえに、さまざまな社会的問題や倫理的問題も生み出している[5)]。たとえば、今日では自動運転の開発が進められているが、そこでは、自動運転車はどの程度の安全性を実現すべきなのか、それはどのようにすれば実現できるのか、自動運転車が事故を起こした場合にそれは誰の責任になるのかといったことが問題となっている。

また、今日では多くの場面でビッグデータを活用したアルゴリズムが利用されているが、そこにはさまざまなバイアスが存在することも問題となっている。たとえば、女性やマイノリティ集団に属する人々が十分な教育を受ける環境が整っていない社会では、過去のデータで学習したアルゴリズムを用いると、それらの人々が採用において不利な判定を下されるかもしれない。今日の人工知能で用いられているアルゴリズムにはどのようなバイアスが含まれており、どうすればそれを解消できるかということも、現在重要な問題となっている[6)]。

このことと関連して、深層ニューラルネットワークの透明性もしばしば問題となる。さきにも述べたように、深層ニューラルネットワークはきわめて複雑なモデルであり、その働きには未知の部分が多くある。それゆえ、ある特定の事例においてあるネットワークがなぜある出力を生成したのかということを、人間に理解できる仕方で説明することはしばしば困難である。われわれにとって重要な決定に深層ニューラルネットワークを用いるときには、このことは深刻な問題を引き起こす可能性がある[7)]。

---

4)　深層学習に関する理論的な研究では、現在これらの問題が活発に検討されている。これについては第 3 章を参照。

5)　人工知能をめぐる倫理的問題に関しては、たとえばクーケルバーグ（2021）を参照。

6)　アルゴリズムバイアスに関しては、たとえばオニール（2018）を参照。

7)　この問題に対処するために、説明可能な人工知能（explainable AI: XAI）の研究開発が近年盛んに試みられている。その具体的な内容については、たとえば大坪ら（2021）を参照。

人工知能の社会的影響をめぐっては、より長期的な問題も論じられている。たとえば、これまで人間が行ってきた仕事が人工知能に代替されることで、人間の職の多くが失われるという可能性が近年話題となっている。楽観的な論者は、これによって人類は歴史上はじめて労働から解放されることになると、この事態を肯定的に評価する。しかし、悲観的な論者は、これによってわれわれの多くが失業し、収入源を失うことになると、この事態を警戒している。産業構造の変化は、産業革命期などにも生じたことである。しかし、人工知能は人間が行っているほぼすべての仕事を代替できるかもしれないという点で、現在われわれが直面している状況は、これまでに経験した状況とは根本的に異なる可能性がある[8)]。

さらに長期的な問題としては、人工超知能（artificial super intelligence）の誕生も議論されている。将棋や囲碁といったゲームにおける人工知能の強さを考えれば、将来、われわれよりも優れた汎用知能をもつ人工超知能が出現することも考えられるだろう。そのような人工超知能がわれわれの意に反した振る舞いを見せたとき、あるいは、人間に対して敵対的な態度をとったときには、われわれが人工超知能を制御できるのかどうかが問題となる。悲観的な論者は、そのような事態が生じたときに人間が人工超知能を制御することは不可能なので、人工超知能の開発や、人工超知能の誕生につながるような人工知能研究は行うべきではないと主張している[9)]。

本書の問題設定

このように、過去 10 年ほどの間における人工知能の急速な発展によって、人工知能をめぐっては、短期的で現実的な社会的・倫理的問題と、長期的で理論的な社会的・倫理的問題の両者が活発に議論されている。しかしながら、本書の目的はこれらの問題を論じることではない。本書が目を向けるのは、両者の中間に位置するような問題である。

このような問題に着目する 1 つの理由は、人工知能研究の究極目標と現状の間にあるギャップである。さきにも述べたように、人工知能研究の究極目標は、

8)　この問題については、たとえばブリニョルフソンとマカフィー（2015）を参照。
9)　この問題に関しては、たとえばバラット（2015）を参照。

人間のような汎用知能をもつ自律的エージェントを人工的に作り出すことである。しかし、現在われわれの周りにある人工知能は、非常に高性能ではあるものの、特定の課題に特化したものでしかない。

われわれはこの現状をどう評価すべきだろうか。1つの評価は、人工知能研究は究極目標をまだ達成できていない、というものだろう。このような評価によれば、現在われわれの身の回りにある人工知能は、究極的な人工知能の不完全な実現、あるいは部分的な実現にすぎないものだということになるだろう。

しかし、人工知能研究の現状には、別の評価を与えることも可能だろう。それは、現在の人工知能研究は、当初構想されていたものとは異なる目標を追求しており、そしてその目標を十分に達成している、という評価である。このような評価が可能であるのは、人工知能に関しては2つの異なる見方が可能だからである。

人工知能に関する1つの見方は、人工知能を、汎用性があり自律的な知能をもつもの、言い換えれば、1人の人間に置き換わることが可能であるようなエージェントと見なす見方である。人工知能研究の初期から、その究極目標を定めてきたのは、このような見方にほかならない。以下ではこれを、代替物としての人工知能、あるいは主体としての人工知能と呼ぶことにしよう。このような見方の下では、人工知能研究がその目標を達成するまでには、まだ多くの課題が残されているということになる。

人工知能に関するもう1つの見方は、人工知能を一種の道具と考える見方である。この見方によれば、人工知能に期待されるのは、人間に置き換わることではなく、火や自動車、携帯電話などと同じように、人間の能力を高め、拡張することである。このような見方を、補完物としての人工知能、あるいは道具としての人工知能と呼ぶことにしよう。このような見方によれば、自律的な汎用人工知能の実現は、人工知能研究の唯一の目標ではなく、短期的には、最も重要な目標でさえないということになる。道具としての人工知能を開発するという観点に立てば、人工知能研究は、すでに多くの重要な成果を挙げてきたと言うことができるのである。

人工知能について考える上では、これら2つの見方を明確に区別することが重要である。そして、人間と人工知能の関係について考える上では、当面は、

道具としての人工知能という見方が特に重要だと思われる。では、人工知能を道具として見たときには、そこにはどのような特徴があるのだろうか。われわれの社会をよりよいものにするためには、道具としての人工知能をどのような形で活用すべきだろうか。これが本書の基本的な問題関心である。