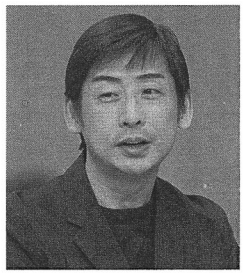
【対談】

脳科学とAIの現在を語る

――『東大塾　脳科学とAI』をめぐって

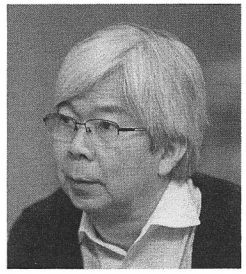


（鈴木貴之）

酒井邦嘉
東京大学大学院総合文化研究科教授

×

鈴木貴之
東京大学大学院総合文化研究科教授



（酒井邦嘉）

編集部　本日は、二〇二四年一一月に小会より出版した『東大塾　脳科学とＡＩ』の編者であり、本書のもとになっている「グレーター東大塾」（一〇人の講師によるリレー講義）の塾長を務められた酒井邦嘉先生と、本書の執筆者であり、グレーター東大塾の副塾長を務められた鈴木貴之先生に対談していただきます。

はじめに

酒井　グレーター東大塾をオンライン講座として開催したのは、コロナ禍の二〇二一年秋期でした。それから三年が経ちましたが、特に昨年からの「生成ＡＩ」と呼ばれる技術の普及は、大きな変化の一つだと感じています。講座の当時、ＡＩが人間にとって脅威となるという可能性はそれほど明確ではありませんでした。

今や世間は空前のＡＩブームに沸いている一方で、ＡＩが人間の脳にもたらす危険性や懸念は深刻化しています。本書が出版されたタイミングでそのことに一言触れなくてはと思い、あとがきを書きました。「楽をしたい」という理由からＡＩを使えば、「効率化」と

という免罪符が得られる風潮です。ところが、その代償として脳を使う思考や創造が軽んじられるわけで、これは特に教育にとって危機的な状況なのです。

大学の学生たちは、文章を書くという行為そのものをしなくなってしまうかもしれません。実際、昨年から私の講義では、レポート課題をすべて廃止し、筆記試験に変更せざるを得ませんでした。学生にはその理由を説明しましたが、ＡＩをどの程度使用したかが教員側でわからない以上、お互いに信頼関係が損なわれてしまう事態を避けたかったのです。

学生がＡＩの使用を正直に申告するかという倫理性や、虚偽申告があった場合の対応といった新たな問題も生じています。結果として、ＡＩや電子機器が一切使えない筆記試験や口述試験の場でしか、学生の到達度や実力を判定できなくなりました。これは将棋や囲碁でも同じで、実際の対局ではＡＩを使わないことが前提なのです。普段からＡＩに頼り切っている人が本番で使えなくなれば、結果は推して知るべしでしょう。

本書の一〇講は脳科学やＡＩ技術の紹介だけでなく、倫理的な観点からも問題を掘り下げています。

鈴木　授業でＡＩをどのように扱うかは非常に難しい問題です。私は筆記試験と小レポートの両方を課してきましたが、長文レポートについては、ＡＩの使用を認めています。その際には、ＡＩを使った場合、減点はしないのでどのように活用したのかを具体的に記載するよう指示しています。たとえば、優れた学生の中には、「このプロンプトを使用した結果、このような回答が得られた」と詳細に記載してくれる人もいます。しかし、哲学のレポートでは、ChatGPT を利用しても、Wikipedia などの情報を単にまとめただけの内容にとどまってしまうケースが多く見受けられます。

また、自己申告はしていないものの、実際には生成ＡＩを利用している学生も一定数いるのではないかと感じています。こうした状況から、これまでのように学期末レポートを書かせること自体が難しくなってきています。私も今学期からは、試験時間内に記述を行わせる形式に変更せざるを得ないと考え始めています。

酒井　先日、『デジタル脳クライシス』（朝日新書、二〇二四年）という本を出したのですが、ＡＩに限らずデジタル機器の過度な使用が、人々の心や対話を大きくゆがめています。

特に哲学においては対話が基本であり、それがすべての学問の伝統として続いてきました。この伝統は古代ギリシャ時代から受け継がれており、学問の基礎に根付いています。しかし、相手の意図を理解しない「対話風」のＡＩとの間に、有益な対話などできるはずはありません。決して資料検索の域を出るわけではないのです。

鈴木　検索エンジンや Wikipedia を利用する学生は以前からいましたが、最近では ChatGPT を使って情報を得る学生が増えています。東大でもそのような動きが顕著になっています。たとえば、授業の資料作成時にもChatGPTに質問する学生が増えています。彼らはとりあえず生成ＡＩに尋ねれば役立つ情報が得られると考えているようです。

酒井　最近は、検索エンジンのトップにＡＩによる要約が表示される仕様になっています。そうした「要約」があたかも代表的な意見であるかのように思ったり、無批判に受け入れたりする危険性が高くなりました。多くの人は検索結果の上位しか見ないですし、そこにＡＩが介入することで、影響がより深刻化します。ネットの情報空間が加速的にゆがみ始めているのです。

また、文書の作成や表示ソフトでも、ＡＩによる要約や「たたき台」の作成を無条件に誘導する仕組みが追加されつつあり、字が書けなくなるばかりか、文章を作れなくなる方向へと強制的に追いやられています。

そうした新たに追加される余分な「学習」機能は、無効化したくても限界があるものです。このように各企業が競うようにＡＩを取り入れている現在の風潮は、知的作業を明らかに退化させる効果を生んでいます。

さて、話を本書に戻しますが、鈴木先生は講義に参加されていたときの印象はいかがでしたか。

鈴木　講義の段階から「脳科学とＡＩ」がテーマでしたが、ＡＩに関する話題の比重が非常に大きかったように思います。脳科学の議論をする中でも、ＡＩが占めるウェイトがますます大きくなり、無視できない存在となっていると感じました。講義の段階からその点は強く意識していました。

「何かのための脳」という誤解

酒井　本書のあとがきで書きそびれたのですが、各講義後の質疑応答（Ｑ＆Ａ）は、編集部のほうで典型的なものを二、三選んでいます。実際の講義では討論がとても活発で、一八：〇〇―二〇：三〇という十分な時間枠を確保していながら、いつも一時間を超す白熱した討論が行われていました。

また、本書では講義で用いられたスライドの一部を再録しつつ、各講師が内容について現時点における最新の知見で改訂してくださいました。編者として私が手入れをした部分は特にありませんが、成書として統一すべきか悩んだところはあります。

たとえば一九二ページの「何のための脳?」という部分ですが、三二ページの「生物進化には、そもそも『目的』や『必要性』はありません。人類がさらに賢くなるべき目的や必

要性があったとしても、そのために脳が進化することなど科学的にありえないのです」、と私が述べた説明と明らかに矛盾します。

ノーム・チョムスキー（アメリカの言語学者）は、“That's why it makes no sense to say that some system evolved for X (‘the spine evolved for keeping us upright,’ or ‘language evolved for communication’).”（Gengo Kenkyu, 2021, p. 11）と明快に述べています【あるシステムが「Xのために」進化したと言うのは無意味なのである（「直立を維持するために背骨が進化した」「コミュニケーションのために言語が進化した」等）】。

ですから、「人の脳は社会性や心を読むという機能のためにサイズが大きくなった」「人間の脳が大きくなったのは非常に栄養価の高いものとしてフルーツを見つけるためだ」「脳が存在する本当の理由は、体を動かすためだ」「精密・精巧な運動をするために脳は進化したのだ」といった専門家たちのもっともらしく聞こえる主張（一九二―一九三ページ）は、すべて誤りです。実際には、人間の脳が進化した結果として、「精密・精巧な運動」などができるようになっただけのことです。

こうした根強い誤解が生物に関するテレビ番組などで連呼されている背景には、生物学者や脳科学者たちの空虚な議論があるわけで、そうした学問の現状も同時に伝えるべきだと考えた次第です。

鈴木　たしかに、「何かのために存在する」という表現にはミスリーディングなところがあります。たとえば、「脳のサイズが大きい個体のほうが社会集団をうまく維持でき、生き残った」といった事実があった場合、それを「社会集団を維持するために大きな脳が存在する」としてしまうのは誤りです。一般向けの説明では、このように単純化された表現が多く見られます。

本来、進化のプロセスには目的はなく、それを比喩的に単純化した表現で説明しているのだということを踏まえなければなりません。この点に注意しないと、誤解を招く危険性があります。

酒井　それは、分かりやすさのために科学的な正確さを犠牲にした典型例でしょう。

『スマホ脳』（新潮新書、二〇二〇）の著者であるアンデシュ・ハンセンはスウェーデンの精神科医ですが、たとえば「感情というのはもともと、キリンの長い首やシロクマの白い毛皮と同じように、生き延びるための戦略だった」（同三六ページ）と決めつけているほどで、このほかにも科学的根拠の乏しい意見が散見されます。

今回の本のように科学の最先端をそのまま提示しようとすれば、研究者の見解が科学的な真理と食い違うこともあり得たわけです。

鈴木　その点については、注釈などを補った方がよかったかもしれませんね。

酒井　人間の脳がどのように進化してきたかについては結果論しかないのですから、想像をたくましくし過ぎてはいけません。

たとえば、現生人類はネアンデルタール人よりも脳が一割ほど小さくなったことが知られていますが、「言語を獲得することで効率的に情報を処理できるようになって、すこしですが脳が小さくなっているのではないか」（本書一九二ページ）というのは、おかしな議論です。それに、化石人類に言語が誕生したのは、ネアンデルタール人より何万年もさかのぼる可能性があります。

言語は進化において中立的なものであり、

生存環境への適応の産物ではないと私は考えています。この点についてはいかがでしょうか。

鈴木　言語を使用するメカニズムを維持するコストに見合ったメリットがあったのだとは思いますが、元からある何らかの目的のために進化したのではないという意味では、その通りだと思います。

酒井　フィリップ・リーバーマン（アメリカの人類学者）は、現生人類では喉頭の位置が下がって誤嚥のリスクが増したにもかかわらず、母音を発声できるというメリットが言語を誕生させたという説を述べ、一般にも流布しました。しかし、手話の使用という言語能力を考えれば、喉頭の位置など無関係ということになります。要するに脳が発達した結果として、完璧な設計の言語が生まれたというだけの話です。

科学的に実証できない「言語の起原」を昆虫や鳥など他の生物種に求めるような説も、生存環境への適応を前提としているという点ですべて間違いなのです。言語の際立った特徴である「可能無限性」（『きりなしうた』や『つみあげうた』のような無限の生成能力）は、有限の変化を積み上げても到達できないものですから。

メタ的な思考とコミュニケーション

酒井　哲学の分野では、言語についてどのように扱っているのでしょうか。

鈴木　哲学でよく話題になるのは、言語の本質的な役割は、個体間のコミュニケーションなのか、個体内の思考なのかということです。言語は個体間のコミュニケーションのためにあるという考えは自然ですが、哲学では、自分自身との自己対話や思考においても言語が役立つという点も重要視されています。たとえば、ダニエル・デネット（アメリカの哲学者）がこのようなアイデアを強調していることは有名です。デネットは、脳の中にある情報を言語という形で一旦外に出すことで、ただ頭の中で情報を処理するのとは異なる新たな情報処理のプロセスが可能になるとしています。デネットは、これは非常に大きな飛躍だと考えています。これは、言語の

役割をめぐる重要な議論の一つです。

酒井　「コミュニケーションのために言語が進化した」という例を先ほど挙げましたが、これが誤りなのは目的論だけでなく、「言語＝コミュニケーション」という見方それ自体にもあります。コミュニケーションは「外言」にすぎず、言語による理解はすべて思考言語を含む「内言」によって支えられているのです。

　この内言とは独立して、脳の感覚運動系というインターフェースを介することで外言が生じるだけであって、そこに「新たな情報処理のプロセスが可能になる」と言うのは本末転倒だろうと私は考えます。

　一方、「自分自身との自己対話」というのは、最も人間らしい言語能力の表れだと言えるでしょう。ほかの動物では無理ですからね。本書でも取り上げられている「再帰性」は、自分が考えたことを自分で意識できるという性質や、哲学というメタ的な思考の根幹であります。言語学もまた、言葉に対するメタ的な分析や説明を言葉で行うわけです。

鈴木　哲学では、言語なしにメタ的な思考が可能かどうかについても議論があります。とはいえ、言語を持つことで、自分が何を考えているのか、自分が何をしたいのかをはっきりと把握できるようになります。その結果、より戦略的な行動が可能となり、場合によっては他者を騙すといった行動も巧みに行えるようになります。この点は、哲学でも注目されるテーマです。

酒井　昆虫の擬態や、鳥の擬傷行動も、われわれには騙しの一種のように見えるかもしれませんが、それらは単に遺伝的に決定された形態や行動によるものであって、「戦略的な行動」ではありません。言葉を持たない動物がメタ的な思考を持つことはないでしょう。

鈴木　そもそも動物行動学では、実験を通じてチンパンジーなどがどの程度高度な「騙し」の行動を行うのかが研究されています。このテーマについては哲学者も多く議論しており、それをどのように解釈すべきかがしばしば論じられます。しかし、明確な結論を出すのは難しいです。どんなに巧妙な実験を行ったとしても、結局は「複雑な条件付け」によって行動しているだけであり、相手の考えを読んで行動しているわけではない、という説明が可能です。そのため、実験だけでこの問題に明確な結論を出すことは難しいわけです。

酒井　その通りだと思います。「賢いハンス」（計算ができると思われた馬）の逸話と同じことですね。

　ゴードン・ギャラップ（アメリカの心理学者）が鏡を用いて行った自己認識の古典的な実験では、チンパンジーがパスしたのに、サルはパスしませんでした。一部の類人猿が「鏡に自分が映っている」という認識を持つとしても、メタ的な思考を持つにほど遠いわけですが。

ロボット三原則をめぐって

酒井　その一方で、ロボットにどの程度の自己認識を持たせて、それ自体を律するべきかは、避けて通れない課題だと思います。アイザック・アシモフ（アメリカのSF作家）は、次のような「ロボット三原則」を提唱したことで有名です。

> 第一条　ロボットは人間に危害を加えてはならない。また、その危険を看過することによって、人間に危害を及ぼしてはならない。
>
> 第二条　ロボットは人間にあたえられた命令に服従しなければならない。ただし、あたえられた命令が、第一条に反する場合は、この限りではない。
>
> 第三条　ロボットは、前掲第一条および第二条に反するおそれのないかぎり、自己をまもらなければならない。

　ところが、現実のロボット工学ではこれらの原則を一顧だにしていないのは、なぜなのでしょう。たとえば、人的被害を回避しえない自律型のロボット兵器や、自爆攻撃兵器は明らかに第一条と第三条に抵触します。それでいて第二条のような命令の絶対的服従については全く譲らないわけです。ロボットの平和利用を無視してパンドラの箱を開けてしまった人類に、未来はあるのでしょうか。

鈴木　私はあまりSFに詳しくはないのですが、ロボット兵器など、政治的理由によって第一条のような原則が無視されるような場合だけではなく、自動運転車のように兵器とは異なる場面でも、この原則を実装することは難しいように思われます。現実世界には例外的な状況が数多く存在し、それにどう対応するのかが非常に難しい問題となるからです。

　古典的なAIでも同様の問題がありました。ルールベースのAIでは、杓子定規にルールを適用することで不適切な結果が生じることが頻繁に起こりました。人間の場合、常識によって「これは例外的な状況だからルールを適用しない」と判断できますが、AIにはそのような柔軟性がありません。明示的に例外状況を定義しなければ対応できないというのが、古典的なAIの大きな問題でした。この課題は自動運転車など現代のAIでも問題となるものであり、非常に重要です。

酒井　自動運転技術では、常に事故の可能性を監視して予防措置を講じることが求められるわけで、「第一条」という形で明示されなくても、ドライバーや同乗者の安全を守るのは当然です。しかし実際の運用においては、交通事故の可能性がある場面でジレンマが生

じます。たとえば、対向車が車線を越えて向かってきた場合、急ハンドルを切れるのかといった問題が発生しますし、それによって後続の車を危険にさらすかもしれません。こうした状況は、いわゆる「トロッコ問題」のような倫理的ジレンマを常に伴うものであり、メーカー内や技術者間の議論だけでは不十分でしょう。倫理的な視点も含めて対処すべき重要な問題だと思います。

鈴木　確かに、自動運転に関連する議論では、トロッコ問題のような究極の選択が話題になることが多いです。しかし、これは人間であっても正解がわからない問題なので、ＡＩにとっても困難であるのは当然です。

　むしろ、人間ならばある程度柔軟に対応できる状況でさえ、自動運転車が現状ではうまく対応できないということが、より大きな問題だと考えます。たとえば、ある選択肢では死者が出るが、別の選択肢では軽傷者が出るだけで済む場合、人間であれば瞬時に「軽傷者が出る方がまだ良い」という判断を下せることがあります。しかし、ＡＩの場合、危害の定義や状況に応じた判断基準を事前に厳密に設定しなければ対応できません。この点が現実的な課題であり、ＡＩに人間と同程度の柔軟性を持たせるには、まだ多くの課題が残されています。

酒井　確かに、ＡＩが人々のデータベースを利用しているからと言って、人間にとっての常識がＡＩに通用すると思ったら大間違いですね。人が道路の横断歩道を渡るということすら、ＡＩにとっては想定外になり得るかもしれません。

　昨年七月にモンゴルのウランバートルを訪れたのですが、信号機のない横断歩道を渡るのに難儀しました。両方向からの車が絶えることがなく、横断歩道の前に立って待っていても速度を緩めてくれません。ところが、反対側におばあさんが現れて横断を始めたとたん、「モーゼの海割り」のように両方向の車両が止まったのです。モンゴルの運転手は、どうやら歩行者の動きを見て判断しているようです。そうした「阿吽の呼吸」は、人間の文化や習慣の現れなのでしょう。

鈴木　国や地域によって人々の歩行や動き方、信号を無視する頻度などが大きく異なるのは確かです。

酒井　私の経験した範囲ですが、横断歩道の前に立つだけでいつも車が止まってくれるのは、ボストンと松本市だけでした。自動運転には、運転技術だけでなく運転マナーも必要でしょう。

鈴木　交通ルールをただ遵守するだけでよいというわけではありません。歩行者がルールを無視するような環境においても、それを計算に入れて対応する必要があります。これが非常に難しい課題です。

酒井　そうした人間の行動の予測となると、さらに膨大な計算量が必要になりますから、ＡＩの「フレーム問題」を再燃させる可能性が高いと言えます。

　私が特に関心を持っているのは、「人間の範疇」という問題です。第一条の「ロボットは人間に危害を加えてはならない」にある「人間」、そして第二条の「ロボットは人間に与えられた命令に服従しなければならない」とある「人間」とは、どのような人間を指すのでしょうか。その命令を出した人が独裁者やテロリストだった場合、ロボットはどこまでその狡猾さや残忍性を見抜けるでしょうか。

鈴木　確かに、そのような問題は非常に難しいものです。ただ、人間の場合、ある程度の柔軟性を持って対応することができます。たとえば、命令を出したのがテロリストであれば、「その命令は例外だ」と判断することができます。また、法に反する命令を受けた場合、それに従うべきかどうかを判断する場面もあります。こうした場合でも、人間はそれなりに適切な判断を下すことができます。ただ、それがなぜ可能なのか、どうして私たちがそのような能力を持っているのかについては、いまだに解明されていません。

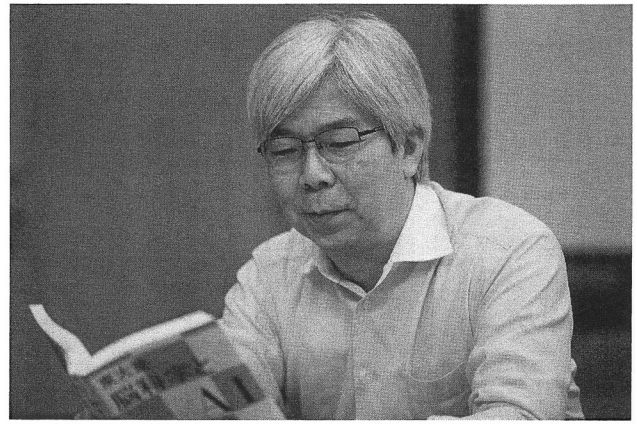
　ヒューバート・ドレイファス（アメリカの哲学者）は、こうした問題に関して鋭いＡＩ批判を行ったことで知られています。彼は、「ルールだけでは人間の柔軟な判断を捉えることはできない」と主張しました。たとえば、横断歩道のような状況や、命令に従うかどうかの判断も、単純なルールでは捉えきれないというのが彼の考え方です。ドレイファスは、人間が文化の中で育ち、経験を積むことによって柔軟な対応が可能になると述べています。この主張は、おそらく事実だと思います。長年の経験や文化的背景の中で蓄積された知識が、人間に柔軟性をもたらしているのでしょう。

　ただし、それが具体的にどのように機能しているのか、また同じことをロボットに実装しようとした場合、どのようなメカニズムが必要なのかについては、まだ十分に解明されていません。

ＡＩ活用におけるルールデザイン

酒井　人間に本性として生得的に備わる言語能力については、理論言語学や言語脳科学によって解明が進んでいます。しかし、経験値については、まだ脳科学の研究テーマになっていないのです。人間自身が人間を科学的に理解できていないこの状況下で、ＡＩをデザインするというのは、かなり乱暴な行為ではないでしょうか。

　自分の素性を隠して不特定多数に向け発信できてしまうＳＮＳにしても、人々の悪意や悪用を想定せずにデザインされてしまい、その付けが回って来ています。結局、「人間とは何か」という問いに十分向き合えず、人間



興味深いです。たとえば、画像認識や将棋の対局のように、ＡＩが優れた能力を発揮する場面は、非常に限定的で、ルールが明確な環境です。このように、イレギュラーな事態が発生しない範囲では、ＡＩは問題なく動作します。しかし、現実世界ではそうした枠を超えた複雑な問題が頻繁に発生します。

たとえば、人命のために交通違反が必要になる場面があります。これは非常に複雑で総合的な判断を必要とする問題であり、単純なルールや計算能力では対処できません。すべての状況を勘案し、最終的に「これが最善だ」と結論づけるような仕組みが必要です。こうした課題は、将棋や画像認識よりもはるかに複雑で、多くの要因が絡み合うため、現状のＡＩでは対応が困難です。

酒井　そのように考えてくると、環境や領域を限定せずに万能であるかのように働かせようとする「汎用ＡＩ（artificial general intelligence, AGI）」ほど恐ろしいものはないでしょう。人間側の強い期待感からか、そもそも文書合成ソフトにすぎない「生成ＡＩ」が、万能の知恵袋のように扱われている現状を考えれば、ＡＧＩのリスク管理ができないまま、人類はＡＧＩによって滅ぼされることになりそうです。

実際、日本の各自治体がＡＩを導入して、公共サービスの効率化を図ろうとしています。対話風ＡＩの答えが市政運営を左右するかもしれないのです。企業や自治体が効率や経費削減を優先してＡＩを導入する一方、人々の価値や要請を正当に考慮しない状況は非常に深刻です。「市民を人間として見ているのか」という疑問を抱かざるを得ません。

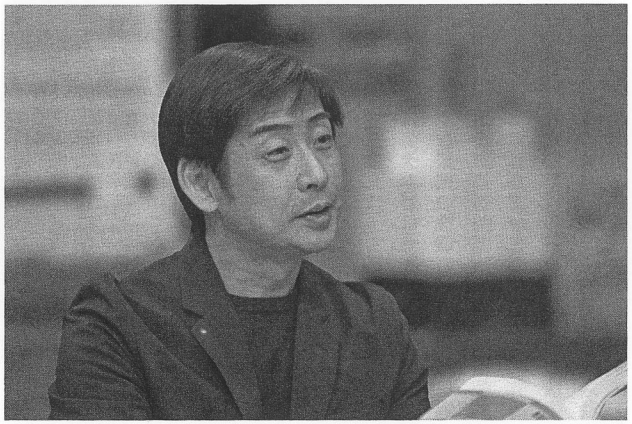
鈴木　一方で、日本のように労働力不足が深刻な場合には、ＡＩをうまく活用する必要があるという見方もあります。

酒井　しかし、企業の切迫した経済的状況からすれば、「ＡＩを活用する」という目的が人員削減を正当化する理由にもなってしまうでしょう。

ロボット兵器を開発する口実も同様です。兵力の不足を補い、「兵士の命を守るため」という大義名分のもと、ＡＩを活用した自律型のロボット兵器が各国で増産されています。大量破壊兵器と同じで、ロボット兵器の保持による抑止力は幻想でしかなく、あらゆる種類の軍備が際限なく拡張することになり

性の問題を十分に理解できないまま放置してきた結果、これらの問題が解決されないまま悪循環を生んでいるようです。

鈴木　今おっしゃったような問題と現在ＡＩが威力を発揮している分野の対比は、非常に



ます。このような状況で犠牲になるのは兵士ではなく、一般市民ではありませんか。

仮にＡＩによって重大な事故が発生したとして、誰がその責任を取るのでしょう。ＡＩメーカー、ＡＩの開発者、ＡＩの導入を決定した人たちは、いずれも被害を予見できなかったと主張するでしょうから、結局はＡＩを使用した人の自己責任とされる可能性が高いわけです。「新技術を恐れるな」「ＡＩを規制しても何も始まらない」と声高に言う人たちは、技術開発の名のもとにリスクを無視するだけでなく、人間性や脳の能力をも軽視しているのだと思います。

合理論と対立するＡＩの経験論

酒井　さて、そうした人間を軽視するようなＡＩの思想的背景には、「経験論」があります。経験論では、人は生まれつき「タブラ・ラサ（白紙の状態）」で生まれ、経験によって何でも書き込めるとしています。この極端な立場では、「機械でも人間と同じように何でも学習できる」という前提につながり、現在のＡＩの基本理念となってしまったわけです。

ジェフリー・ヒントン（カナダの計算科学者）は、「人間には言語を使う機能が生まれつき備わっているという主張もあるが、それは全くのナンセンスだ。言語は生まれた後に学習する後天的なものだと考えている」と主張しています。このように経験論至上主義のＡＩ開発が技術を誤った方向へ誘導しているわけですが、そうした人間軽視の技術が先行する限り、人類に未来はないと思います。

先ほど紹介したチョムスキーは、デカルト由来の合理論の立場から言語の生得的な性質を明らかにしてきましたし、私もこの言語理論に基づいた脳科学の研究を三〇年ほど続けてきました。つまり脳科学とＡＩの背景には、合理論と経験論の激しい対立があるわけです。哲学の現在地はいかがでしょうか。

鈴木　一般的に言えば、合理論と経験論の対立については、特に人間の心に関しては、ある種の折衷案的な立場が取られていると思います。人間の心に進化的な基盤があることは当然認められるべきで、極端な経験論は無理がありますが、他方で学習も重要であるということで、両方の要素を組み合わせる考え方が一般的になっているように思います。

また、ＡＩとの関係について言えば、古典的なＡＩは合理論的な側面が強いですね。ルールベースで設計され、必要な知識を最初に与えてそこからスタートするという仕組みです。一方で、ニューラルネットワークは経験

論的なアプローチを取っています。データから学習し、その結果として何が得られるかが決まるという形です。このように、AIにおいても合理論と経験論という二つの視点が対比される形で考えられていると言えるでしょう。

酒井　一九七〇年代に開発されたAIの「エキスパートシステム」は、知識と経験をあらかじめ組み込んだ上で合理的な判断を追求するもので、論理計算を実現するという意義がありました。たとえば「マイシン」というエキスパートシステムでは、医学的判断の基礎となるルールや薬剤の性能をもとに、細菌感染症の患者データに対して抗生物質の治療方針を提案することができました。

　一方、一九八〇年代のPDP（並列分散処理）モデルを境に、現在の機械学習は経験論を温床として、前提知識が全くない状態からスタートさせルールを一切廃するという、ゆがんだ形で増幅してしまいました。そこで生じた言語学とAIの間のあつれきは、合理論と経験論の対立構図の中で一層広がり、根深いものになってしまいました。引き戻すべき方向性があるとすれば、脳科学の合理論的なアプローチに戻ることではないでしょうか。

鈴木　私は、ある意味で、それとは逆のところに期待している部分が少しあります。非常に大規模な深層ニューラルネットワークは、とにかく大量のデータから学習させると、「何がどうなっているのかはよくわからないけど、すごいパフォーマンスを発揮する」というのが現状だと思います。現時点では、基本的なメカニズムについてはわかっていますが、より大規模なスケールでどのような情報処理がなされているのかは、あまり理解されていません。

　もしそこを数理的にきちんと説明できるような理論やツールが出てくれば、従来の合理論的なアプローチ、いわゆるルールベースのアプローチとは違った形で、知能をある程度抽象的に記述できる新しい理論が得られるのではないかと思います。それが実現できたら、すごく興味深いですね。

酒井　将棋や囲碁のAIでも、おそらくそれは無理筋でしょう。山本一成（いっせい）さんの『人工知能はどのようにして「名人」を超えたのか？』（ダイヤモンド社、二〇一七）によれば、ディープラーニング（深層学習）はもはや科学と訣別してしまっていますから。

鈴木　そもそも哲学者は、「私たちが普段使っている日本語や英語のボキャブラリーで心のメカニズムを語ることができる」という前提を持っています。つまり、今われわれがもっている概念的な道具立てだけを使って心について語れなければ、哲学として成り立たないという考え方が基本にあるわけです。

　しかし、それとはまったく異なる方向性、例えばベクトルや関数といった数理的なツールを使うことで、心のメカニズムを記述することが可能だという考え方が、もし具体的な形で示されるようになれば、それはとても興味深い展開になると思います。ただ、現時点では、従来のオーソドックスな理論に代わる具体的な代替案が何なのか、まったく明確になっていません。そもそも、代替案が本当に存在するのかどうかさえも、まだ分からないという状況です。

酒井　機械学習とは完全に独立した形で、数理的なツールを使う方向性は常に残されていると思います。脳に対する入出力に近い処理、すなわち知覚や運動のメカニズムであれば、特徴抽出や運動制御などの理論として、

数理モデルが大いに役立ってきました。脳の高次機能となると、記憶までは数理モデルで解明できそうですが、感情や意識、さらには判断・思考・自意識となると、モデルの可能性を絞り込むのが困難です。そのため、心の要素をAIに取り込もうとすると、状況はますます混乱して方向性を見失いかねません。

　一方、理論言語学では、一九八〇年代から「最小性」という考え方によってモデルを絞り込むことに成功しています。ルールの存在を全否定するような経験論の立場では、言語現象を説明することはできないのです。

鈴木　言語に関しては、昔から似たような議論がありましたね。たとえば、大規模言語モデルが本当に人間と同じように言語を使えるようになった場合、それが完全にルールなしで実現しているのか、そうではないのか、という問題です。そのような状況になった際、結局のところ、生成文法のような何らかのルールがモデルの内部に表現されている可能性があります。ニューラルネットはそのルールの実装形態（インプリメンテーション）に過ぎない、という見方です。

　これは昔、チョムスキーやジェリー・フォーダー（計算論を支持するアメリカの哲学者）がPDPの研究者と論争していたテーマにほかなりません。フォーダーは「結局、二つの記述は両立可能であり、かつ、計算論的な記述がより本質的なレベルだ」と主張していました。こうした議論が、現在再び蒸し返されているようにも感じます。

酒井　言語の獲得においても、それを一般的な学習や認知発達に含めようとするジャン・ピアジェ（スイスの心理学者）がチョムスキーと鋭く対立して、今なお尾を引いています。

鈴木　最終的にはまた同じ結論に落ち着くのかもしれませんが、そうでないとすればどうなるのかは興味深いです。

酒井　たとえば画像認識でも、どのような特徴が実際に効いていて画像認識が可能になっているのか、その背後にあるルールは何なのか、といったことは全く明らかになっていません。この状況で、「何らかのルールがモデルの内部に表現されている」と見なすのは適

切でないと思います。

哲学と言語学の源流

酒井　もし言語の生得性を否定するなら、三、四歳の子どもたちがなぜ完璧に言葉を話せるようになるのか、という明らかな観察事実を説明できなくなります。それを学習で説明しようとしても、どの言語にも見られる時制や活用変化、さらには格・性・数の一致や機能語といった抽象概念を自力で習得することなど、その段階で発達している認知能力では不可能です。

知識を教わった経験がない人でも確かな能力を持つという、この一見不思議な事実は、ソクラテスの教えとしてプラトンが二千四百年も前に指摘しており、「プラトンの問題」と呼ばれます。そのプラトンからデカルトやフンボルトへ、そしてラッセルとチョムスキーと受け継がれた合理論の系譜があり、そこに学問としての哲学と言語学の源流を見ることができます。

プラトンの議論は、どの程度現代的な意味を持って哲学で議論されていますか。

鈴木　プラトンの議論そのものが取り上げられることは少ないかもしれません。むしろ、プラトンの対話篇においてソクラテスが行っていたこと、つまり、正義とは何か、勇気とは何かといった哲学的に重要な概念を定義しようという試みに注目することが多いのではないでしょうか。プラトンの対話編でも、これらの定義はなかなかうまくいきませんが。

酒井　哲学の出発点としての重要性は揺るがないのですね。

鈴木　その通りですね。ある意味では、現代の哲学者も基本的に同じような問題に取り組んでいると言えます。この意味では、哲学の核心的な課題自体は時代を超えて一貫しているのではないでしょうか。

酒井　先日、「駒場の教養を問う――三〇年後のよりよき世界へ」というシンポジウムに私が参加した際に、プラトンの「洞窟の比喩」を取り上げてみました（『教養学部報』第六五九号に要旨掲載）。この比喩は、洞窟内に捕らわれた囚人が見聞きする世界はゆがんだ現実であり、その洞窟から解放されて外に出た人のみが真実「イデア」を知り得て哲学者になれるということを述べています。現代に当てはめると、囚人はSNSやAIに踊らされた一般人と重なります。彼らが見ている現実がいかにゆがんでいても、それを真実だと信じ込むしかないわけです。

しかし、洞窟から解放されて真実を知った賢人であっても、その真実を囚人たちに伝えることは難しいことでした。明るい世界から戻って暗闇に目が慣れない賢人は、「外に出ておかしくなった」と囚人たちから見なされ、「皆を扇動した」と誤解されてしまいます。ちょうどソクラテスがそうだったように。

先ほどのシンポジウムでは、幸運なことに次の講師が哲学者の國分功一郎さんで、そうした内容は「帰還問題」として議論されていると教えてくれました。

SNSやAIの利用とリスク

酒井　人間の言語の設計は完璧ですが、人間の心は実に未熟です。二一世紀になっても、隣国を攻撃することをやめられないほどです。各国の指導者が「平和」をイデアとすることなく、核兵器やロボット兵器の脅しに頼ろうとします。そこに「抑止力」など存在しないということは、物理学者たちが繰り返し指摘してきた問題でした。

鈴木　第二次世界大戦後の約六〇～七〇年間が比較的安定していたということが、むしろ興味深く思われます。

酒井　しかし、冷戦時代には常に緊張状態が続いていましたし、朝鮮戦争やベトナム戦争もありました。ベルリンの壁は崩壊しましたが、今や極右政党や「自国ファースト」の風潮が台頭して、世界全体が内向きになりつつあります。

鈴木　そういう意味では、第二次世界大戦後に生まれ、日本でずっと平和な環境で暮らしてきた人間にとっては、現在の状況は例外的に感じられますが、長いスパンで見ると、戦争が起きている状況の方がむしろ通常の状態と言えるのかもしれません。

酒井　そこで哲学者や有識者ができることはないのでしょうか。

鈴木　直接的にできることは少ないでしょうが、今日の話題との関連で言えば、たとえば、AIやSNSが社会を悪い方向に変えていく可能性は、関心を集めていると思います。

酒井　実際に、メディア操作がそうした方向に影響を与えていますね。

鈴木　特に選挙などでは、メディアやSNSが非常に大きな影響を与えています。これは間違いないことです。個人的な感覚としても、SNSやAIは、われわれが思っている以上にさまざまな悪影響を社会に与えていると感じます。しかし、それをどうやって客観的に明らかにし、どのように対処すべきかを考えるのは非常に難しい課題です。

酒井　オーストラリアでは、一六歳以下の子どもたちに対してSNSの利用を禁止する法案が可決されました。子どもたちの命にもかかわるような問題が生じていることがその背景にあり、それは他人事ではありません。AIの利用も同様の問題をはらんでいます。ですから、抜本的な法的規制を設けない限り、SNSやAIの使用を適切に制御するのは難しいと思います。

私は、「AIとのつきあい方」という表現に賛同できません。なぜなら、「つきあい方」と言った時点で、AIと付き合うことを前提にしているからです。しかもその表現は、深刻な脅威を何ら考慮に入れていません。「核

兵器とのつきあい方」という言い方が不適切なのと同じことです。

AIを規制する選択肢を持たないとは、リスクをあまりに軽んじています。たとえば日本の人工知能学会は、「一律な利用の禁止は何も生み出しません。積極的に利用する前提で、どのように教育に活用するかを検討すべきと考えます」と述べているほどです。

これまで、核兵器にはバグウォッシュ会議があり、遺伝子組み換え技術にはアシロマ会議がありました。しかし自律型兵器に対する公開質問状を除けば、一般のAIに対して科学者はまだ何もしていないのです。リスク管理が不十分なまま開発が進めば、SFのような最悪のシナリオに直面するのではないでしょうか。

鈴木　今のお話の流れから続けると、私はそこまで悲観的ではありません。反対論者というわけでもなく、AIを使い分けるべきだと考えています。つまり、役立つ場面と注意が必要な場面があるということです。個人的に特に関心があるのは、AI兵器のように明らかに要注意な場面だけでなく、それ以外の場面です。たとえば、人命に直接関わる場面ではなくても、間接的に長期的な悪影響を与える可能性のある場面についてです。

SNSはAIそのものではありませんが、そうした悪影響の一例かもしれません。戦争や人命とは直接関係がないものの、私たちの幸福感や精神衛生にマイナスの影響を与える可能性があると感じています。それだけではなく、AIを導入することで短期的には便利に思えるが、長期的にはマイナスに働く場面がいくつもあるはずです。それが具体的に何なのかを見極めていくことが重要だと思います。

酒井　時間軸を考慮に入れるのは大切ですね。原子力発電などの技術も同様に議論することが必要で、長期的なリスクを見極めるべきだと考えます。

AIの長期的な問題は、思考力や創造力の低下、そして人間性の堕落です。AIを人間の奴隷のように扱うことで、知的な判断を伴う作業までもが軽視され、頭脳労働が蔑視されるようになるでしょう。結局それは空虚な優越感を増長させ、倫理や道徳観をもゆがめます。

実際、対話型AIは、単なる「イエスマン」としてデザインされます。逆に質問したり反論したりするようでは商品にならないからです。面倒な判断や決定はAIの「計算」に押しつけておきながら、自分ではその決定に責任をとろうとしない人が続出するでしょう。子どものうちに学校の作文や宿題のレベルからAIに頼り続けたなら、知的作業への喜びや楽しみを失って無気力になったり、感情の抑制もままならず身勝手になったりする恐れがあります。その兆候はすでに現代人に現れてきているのではないでしょうか。

AIの「適切な」使い方はあるのか

酒井　このように考えてくると、教育の現場にAIを導入するということは、教育自体の自殺行為に等しいのではないでしょうか。自分の頭で考えて新たな知識を得るという創造の過程を放棄させ、機械で置き換えようというのですから。キーボードや自動変換に頼らず自らの手で文字や記号を書き留め、予測変換を使わずに自力で語彙を見つけ、考えが他者に伝わるよう推敲を重ねること。そうした作業の一つひとつがあらゆる学問の基礎にあるはずです。

鈴木　そういう意味では、AIの使い方次第でもう少し改善できる部分はあるのではないかと思います。たとえば、将棋ではプロ棋士がAIを非常にうまく活用しています。プロ棋士は局面をしっかりと理解した上で、AIにもう少し先の局面をシミュレーションさせることで有用な情報を引き出しています。AIが出した手の意味も、プロ棋士であれば後から理解することができます。最終的に、その結果を自分の感覚としてしっかりと取り込むことができれば、AIを有効に活用できるわけです。

しかし、逆に答えだけを単に使うような形では、非常に問題が多いと思います。特に、ChatGPTを利用する人の中には、「とりあえずChatGPTに聞けば正しい情報が得られる」と素朴に信じて使っている人が多い印象です。これはむしろ悪い方向に作用していると感じます。

酒井　将棋や囲碁のAIは、勝つ確率を最大化するような手筋を探索するだけですから、きわめて特殊な「計算」にすぎません。棋士や上級者がAIを研究用に使うからと言って、一般の文書作成や画像合成などにまで敷衍するのは、重大な誤りです。文書に勝ち負けのような絶対的な尺度はなく、内容の評価ともなれば、意味や意図の解析が全くできない現状のAIでは歯が立たないのです。ですから教育にAIを導入すれば、学力低下がすべての教科で連鎖的に広がる恐れがあり、百害あって一利なしでしょう。

将棋や囲碁であっても、一手の価値判断ができない初心者や中級者がAIに手を出すべきではなく、「答えだけを単に使うような形」になるのが関の山でしょう。私はチェスのAIを何度も使ったことがありますが、自分の見落としを指摘されるばかりで、考え方を学んだという手応えは皆無でした。AI相手にテイクバック（待った）を繰り返しても、次でまた間違えれば八方塞がりです。過去の名局を研究したほうがはるかに有用でした。

現状のAIは、バックプロパゲーション（誤差逆伝播法）が登場した八〇年代あたりから完全に脳科学の根拠を失い、短時間で計算さえできれば「何でもあり」のゆがんだ形で増長してきました。データベースの構築に巨額の資金を投じるような現状では、脳科学がサイエンスとしてAIから学ぶべきものはありません。

もともとAIは、言語学や脳科学との接点から生まれたものだったのです。現在のAIブームが過ぎ去り、行き詰まった時にこそ、サイエンスとしてのAIを再構築する機会が巡ってくることでしょう。去年AIに対して授賞されたノーベル賞は時期尚早でした。

鈴木　少なくとも現在のディープラーニングや深層ニューラルネットワークのメカニズムについて、AI研究者がもう少し詳しく解明することができれば、それが何らかのヒントになる可能性があると思います。しかし、現在のようにブラックボックスのままだと、脳科学との関連性を検討する余地がほとんどありません。

酒井　私は脳科学者として、未来のAI研究が動物や人間の脳のデザインに基礎を置くことを願っています。脳にはプリインストールされた生得的な能力があるということを解明し、本当の意味での「学習」とは何か、といった理解を深めるべきなのです。

【編集部との質疑応答】

――「身体性」について、AIや脳科学では

どのように位置づけられていますか。

鈴木　人間と同じように多様なことができる汎用ＡＩを作ろうとする際には、身体性が非常に重要になってくるでしょう。

具体的には、汎用ＡＩを作ろうとすれば、実際にロボットを物理的な形で設計し、現実世界で動き回らせ、自律的に学習できるような仕組みが必要になると思います。現在の多くのＡＩ、特に従来のシステムと人間や生物との大きな違いは、この「身体性」の欠如にあります。

身体性に基づく世界の把握、いわゆるアフォーダンス的な世界認識といったものは、現在のＡＩにはほとんどありません。たとえば、机について考える場合、現在のＡＩはその位置や重量、寸法といったデータを扱います。一方で、人間の場合は、「この机の上に物を置けるか」「持ち上げられるか」といった、自分の運動能力との関わりで対象を捉えています。

現在主流の、いわゆる「デスクトップ型」や「箱型」のＡＩが世界を捉える方法と、生物が世界を捉える方法は根本的に異なっています。そのため、ＡＩと生物ではできることとできないことに大きな違いが生じ、さまざまな場面にその影響が現れると考えています。

私は、人間のように動く身体を持たせ、人間のような汎用知能をもつＡＩを作ることは、原理的には可能だと考えています。人間も最終的には物理的な素材、つまり原子の集まりでできているわけです。ですから、物理的な素材から知的なエージェントを作ること自体は、原理的には可能なはずです。実際に人間や猫のような生物も、そのような物理的な構造から成り立っているからです。

しかし、実際にそれを実現しようとすると、非常に膨大な時間と手間がかかるでしょう。まず、人間のような高度に複雑な身体のメカニズムを持たせる必要があります。そして、学習の蓄積も非常に重要です。生物であれば進化の過程で何億年もかけて蓄積された学習があるわけですが、それに匹敵するメカニズムを人工的にどうやって作り出すかという問題に直面します。

ですから、原理的には可能であるものの、例えば一〇年や五〇年という期間で犬や猫と同じように現実世界で動き回るＡＩエージェントを作るのは非常に難しいのではないかと考えています。

ちなみに、ドレイファスも身体性が重要であると主張しており、「ＡＩを作れるか」という問いは、究極的には「人工的に身体を持ったエージェントを作れるか」という問いなのだという指摘もしています。また、身体を持った人工エージェントであれば、それなりに可能性があると認めているようにも見えます。

ただし、「身体が重要だ」と繰り返し述べてはいるものの、その身体が具体的にどのように知性に作用するのかについては、あまりはっきりした説明をしていません。この点については、むしろ現代の身体性認知科学が補完する部分なのかもしれません。

酒井　身体性という概念を連呼する脳科学者もいることは確かですが、物理学のような厳密科学の観点からすれば、身体性など幻想にすぎません。

人間には脳があることで体を動かす運動指令が可能になりますが、その動作がたとえ不自由であったとしても、その人の脳が充実した知的知能を持ち続けていることに変わりはないのです。「表出できない」という障害を伴うＡＬＳ（筋萎縮性側索硬化症）や植物状態の患者を考えれば、「身体性」を持ち出すことなど、人間の尊厳を無視したあり得ない議論です。脳の問題とは、人格を含めた知的機能の問題であり、「人間とは何か」という根源的な問いでもあるのです。

――ＡＩを学問以外の日常生活やビジネスの場面で活用することについてはいかがですか。打ち合わせのための情報をまとめたりする機能も使われています。

酒井　教育もビジネスも、知的活動の維持と研鑽が必要だという点で同じではありませんか。要約の機会すら失ったら、どんな人間の創造的な能力が残るというのでしょう。思考とは、情報の収集と異なる知的作業なのです。ＡＩやインターネットなどなくても、ニュートンやアインシュタインたちは紙とペンだけで創造的な仕事をしてきました。

――ＡＩに対する法学からのアプローチに期待することはありますか。

鈴木　ＡＩ兵器や自動運転車が何らかの危害を引き起こした場合、その責任を誰が負うのかという問題について新しいアイデアをもたらしてくれることに期待したいです。ＡＩをめぐっては、責任に関する現在の枠組みでは責任の所在がはっきりしなくなってしまう状況が多々あります。この問題は、現在でも活発に議論されています。

ここで、単に「責任を負う人がいない」と結論付けて終わってしまうことは、明らかに受け入れがたいことです。しかし、では具体的にどうすればよいのかということは、よくわかりません。こうした問題に対処する上で、どのような新しい枠組みや考え方があり得るのか、その点に関する貢献を期待したいと思います。

これまでは、個々の人間や企業だけが責任を負う主体となるという考え方が一般的でしたが、それでは対応できない場面が今後ますます増えてくるのではないかと思います。そうした状況に対して、新しい視点からの提案が求められていると感じています。そのあたりの議論は、法学の基礎理論にも関わる話です。何か新しい、面白いアイデアが出てくることを期待したいです。

酒井　自由意志の問題は、脳科学でも未解明です。脳科学では基本的に、「脳から心が生まれる」という一元論の立場をとっています。しかし法学では、心が脳という器質とは独立しているという二元論を貫きます。「私の脳が勝手に指令を出して犯罪行為を引き起こした」と被告人が主張したとして、それを認めるわけにはいきませんからね。

私の研究している言語も、一元論からすれば心がその一部を「言語化」で生み出したものであり、脳が生み出す心と密接に結びついていることを前提としています。もし「神経法学」という分野が成立するなら、自由意志や自己責任という問題を一元論で扱う可能性について、議論する場を設けていただきたいと思います。ＡＩに関する議論が脳科学を逸脱することだけは避けねばなりません。

本書『東大塾　脳科学とＡＩ』を通じて、人間を対象とした脳科学の立脚点を知り、ＡＩという技術の将来について考えることに、重要な意義があると思っています。

（二〇二四年一一月二七日、東京大学出版会会議室にて収録）