生成AIとどう付き合うか

東北大学名誉教授
野家啓一
NOE, Keiichi

物質だっていくらも思考能力を持つことができると説いた
形而上学者たちは、かれらの理性を汚さなかったのである。
ド・ラ・メトリ『人間機械論』

はじめに

2022年11月30日にOpenAI社が公開した生成AI「ChatGPT」の反響はすさまじく、登録者は2か月で一億人を超えたという。もちろん、恩恵とともに人類社会にもたらすリスクも多大であることは容易に想像できる。そのためアメリカやEUではAI開発をめぐる規制強化に乗り出した。アメリカでは2023年10月30日にAI開発に安全性テストを義務づけ、その結果や情報を政府と共有するための大統領令を発出した。EUではより厳しいAI規制法案が年内にも成立予定だと報じられている。これはAIのデータベースに評価システムの登録を求めるもので、違反に対する罰則規定では、最大3,500万ユーロ（約56億円）もしくは総売り上げの7%の罰金が課されるとのことである。

アメリカやEUの迅速な対応に比べると、日本はいささか動きが鈍いように見える。もちろん政府の肝煎りで「AI戦略会議」は立ち上げられたものの、そこでの検討は「事業者ガイドライン」の策定といったレベルにとどまっているようである。現代のようなグローバル社会では、新たな技術に関する知識は一国内や一地域内に押しとどめることはできない。それはすでに冷戦下の核開発競争で経験済みの事柄である。もし技術開発に何らかの歯止めが必要ならば、それは国際的に合意可能なルールに基づかねばならない。大国とGAFAMなど巨大情報産業が手を組んで「国際標準」が出来上がってしまってからでは手遅れなのである。今からでも遅くはないので、日本はG7議長国として国際的なルール作りに積極的に関わるべきであろう[1]。そのために日本学術会議が果たすべき役割は大きい。とりわけ自然科学系と人文社会系の両翼を擁する日本学術会議が貢献できる余地は大

いにありうるはずである。AI問題は技術的課題と同時に倫理的課題の検討が不可欠であり、その点で日本学術会議の積極的な関与を期待したい。

1 人工知能（Artificial Intelligence）略史

Artificial IntelligenceすなわちAIという言葉が初めて使われたのは、1956年に計算機科学者を集めて開催されたアメリカのダートマス会議においてである（日本ではかつて「人工頭脳」という訳語が用いられていた）。その当時の目標は「倫理」ではなく「論理」、すなわち人間の合理的思考（演繹的推論）を高速で遂行する機械を実現することであった。これをAIの「第一次ブーム」という。そこで目指されたのは、人間と同程度の思考能力をもつ機械、すなわちチューリング・テストをクリアするコンピュータの実現である。イギリスの数学者アラン・チューリングは、1950年に論文「計算機械と知能」を発表し、以下のような思考実験を考えた。

> 機械と人間とを別々の密室に入れ、部屋の内部に関して情報をもたない第三者が、テレタイプ装置を使って中にいる人間または機械と交信を行う。さまざまな質疑応答を繰り返した結果、第三者にどちらが機械でどちらが人間であるかの判別がつかなければ、その機械は人間と同等の思考能力をもつと考えてよい。[2]

もちろん「判別がつく」ことと「思考能力をもつ」こととは別の事柄である。だが、生成AIの登場によって高等教育機関の中にはChatGPTを用いたレポート作成を制限した大学もあると聞く。またアメリカの科学雑誌*Science*は2023年1月にChatGPTを使った論文投稿を禁止した。これは提出されたレポートや投稿された科学論文について人間（学生や科学者）が書いたのか、生成AIが書いたのか判別がつかないということを意味する。その意味ではChatGPTはすでにチューリング・テストをパスしていると言ってよい。

続く「第二次ブーム」は、並列分散処理を目指すコンピュータとともに1980年代に出来した。「エキスパート・システム」や「第5世代コンピュータ」という言葉を覚えておられる方も多いであろう。つまり、あらゆる場面に適用可能な汎用AI（強いAI）の実現を目標に掲げるのではなく、医学や法律学や自動翻訳など専門知識に限定しての特化型AI（弱いAI）の活用を目指したのである。とりわけMITのワイゼンバウムが作った「イライザ」は、精神分析医の役目を果たす対話型AIとしてよく知られている。

こうしたプロジェクトはある程度の成功を収めたものの「フレーム問題」という壁に突き当たって頓挫した。フレーム問題とは、形式的にいえば「状態に対して作用するオペレータにより、副作用として何が変化する（あるいはしない）かをどう記述・処理したらよいか」[3]という問題にほかならない。

これは簡単そうに見えて厄介な問題である。たとえば「コーヒーが入ったカップが机の上にある」という状態に「カップを机から口もとに移動する」というオペレータを作用させたとする。その際、カップとともに中味のコーヒーもまた口もとに移動することは言うまでもない。同時にカップの色や形が変化しないことも当然であろう。ところが、第二次ブームのAIにはこうした「常識」をも改めて「フレーム公理」という形で一から教え込まねばならなかったのである。それが膨大な量になることは想像がつく。これ以外にもさまざまな問題を抱えた「第五世代コンピュータ開発」は通商産業省（当時）の後押しを受けて巨費が投じられたものの、社会に実装されることはなく失敗に終わった。

ところが2010年代に入ると、「深層学習（Deep Learning）」の技術を基盤とする生成AIの登場により、フレーム問題はそもそも問題ではなくなった。というのも、ChatGPTはゼロから出発して白紙の上に文章や画像を生成しているわけではなく、インターネット上に溢れている膨大なテキストや画像をデータとして取り込み、それを確率・統計に従って連結し組み合わせることにより、新たな文章や画像を出力しているからである。つまり、既存のデータを収集する段階で、すでにそこには「常識」が埋め込まれており、「常識外れ」の発言はスクリーニングを経ていると言ってよい（もちろん「常識」には偏見や差別表現も含まれており、対処が必要なことは言うまでもない）。それゆえ生成AIの登場により、いよいよ自律的な思考能力（心）をもった「強いAI」が可能になるのではないか、と「第三次ブーム」が起こったゆえんである。

だが、汎用AI（強いAI）の実現可能性については、かなり以前からアメリカの哲学者ジョン・サールによって、「中国語の部屋」と呼ばれる思考実験を通じて否定的議論がなされてきた。これはチューリング・テストをヴァージョンアップしたものと考えればよい。

あなたが部屋の中に閉じ込められているとします。その部屋の中には中国語の記号が入ったバスケットがいくつかあります。ここであなたは、（私と同様に）中国語をひとことも理解することができないとしてみましょう。ところがその部屋には、その中国語の記号を操作するた

めの規則を英語で記した本があなたのために用意されていたことにします。その規則には、中国語の記号の操作がまったく形式的に定めてあります。つまり、中国語の記号の意味論ではなく、統語論に基づいて操作が定めてあります。（中略）さらにまた、プログラマのプログラミングが卓越し、また、あなたの記号操作の能力も卓越しているので、すぐにあなたが送り出す解答が中国語を母語とする人の解答と区別できなくなったと想定します。そこであなたは、部屋の中に閉じ込められ、送り込まれてくる中国語の記号に反応してまた別の中国語の記号をあちらこちらへと動かしたり送り出しているというわけです。[4]

いささか引用が長くなったが、この想定はChatGPTによって実現された「日本語の部屋」と類比的だと考えたからである。サールの考えは「コンピュータは統語論（syntax）は持つけれども意味論（semantics）は持たない」[5]と要約できる。意味論を持たなければ、現実世界との接点を持つことはできない。いわゆる「記号接地（symbol grounding）」の問題である。おそらくChatGPTに尋ねれば、「猫」についての百科事典的知識は山ほど得られるであろう。だがそこには、猫の手触りや鳴き声、臭いや引っ掻かれた時の痛みなど、身体的接触を通じた情報は当然ながら含まれていない。しかし、「猫」の中核的意味は言語的記述によってではなく、そうした身体的次元の接触によって把握されるものであろう。

サールに言わせれば「言語を理解するということ、そしてそもそも心的状態を持つということはたんに、一群の形式的記号を操作するということ以上のものを要求します。すなわちそれは、解釈を必要とし、その記号に付与された意味を持つことを必要とします」[6]というわけである。

2 生成AIは言葉の意味を理解しているか

だが、対話型の生成AIはこちらの意地悪な質問に対してもまともな日本語で答えてくれるし、少なくとも彼我の間で一定の対話は成り立っているように見える。それでは生成AIは言葉の意味を理解していると言ってよいのだろうか。答えは二様に分れる。現今のAI研究を最前線で牽引する松尾豊（東大教授）は、いわゆる記号接地問題に留保を付けながらも肯定的である。

ChatGPTは学習によって、「リンゴ」という単語が「赤い」や「果物」などといった単語と関連が深いというということを理解しているわけです。さらには、ニュー

トンはリンゴが落ちるのを見て万有引力の法則を発見したといわれていること、リンゴはApple社のシンボルであることなども知っています。つまりリンゴがどのような場面でどう使われるのか、社会においてどういう含意をもつのかも知っているわけです。そして、これらの知識と文脈に応じて見事に引き出すことができます。これは少なくとも記号的な意味において、「リンゴとは何か」という概念を形成していることに他なりません。つまり、リンゴの手ざわりや味といった実世界の相互作用を伴なわない範囲においてではありますが、ChatGPTはリンゴという概念を理解しているといえるのではないでしょうか。[7]

つまり、言葉の意味理解をウィトゲンシュタイン流に言葉の適切な「使用（Gebrauch, use）」と捉えるならば、ChatGPTはすでに言語を十分巧みに使用する能力を持っているのだから、言葉の意味（概念）を理解していると言ってかまわない、というわけである。そうした考えに対して、情報学の泰斗西垣通（東大名誉教授）は原理的な観点から異を唱える。

AIが理論的には本来、自律性をもたない他律系だということは、いくら強調してもしすぎることはない。生物とは異なり、自らその作動ルールを内部で創りあげているわけではないのだ。基本的にはコンピュータは指令通りに作動しているだけである。したがって、道徳的な主体などとは無縁であり、AIに自由意思だの責任だのを帰するのは全くの誤りにほかならない。[8]

西垣の議論は明快である。AIは人間が指令したアルゴリズムに従って作動する物質、すなわち機械であり、そこには自律的意志のようなものは存在しない。つまりAIは有用な道具以上のものではありえない、ということである。たしかにAIはロボットなどと結合することによって現実世界に関与する「行為者」とはなりうるが、行為がもたらした結果について責任を問えるような「道徳的な行為者」であることはできない。カント的にいえば、AIは手段として用いうる「物件（モノ）」ではあっても、道徳的主体として尊重されるべき「人格」ではないのである。

だが、AIはすでに社会のさまざまな場面に実装されている。いわばわれわれは「AI＋人間系」とでもいうべき現実を生きているのである。以下では不十分ながら、AI実装社会で問われるべき倫理について考えてみたい。

3 生成AIの偏見と差別

人工知能学会倫理委員会は、2019年12月に「機械学習と公平性に関する声明」を発出した。ここで「機械学習」とはいわゆる「人工知能」のことだが、議論の対象を明確にするため、学会ではこの語を用いている。その際に社会一般と共有したい前提として掲げられているのは次の二点である[9]。

(1) 機械学習は道具にすぎず人間の意思決定を補助するものであること。
(2) 私たちは、公平性に寄与できる機械学習を研究し、社会に貢献できるよう取り組んでいること。

要するに、AIは補助的道具にすぎず、使用に当たっては「公平性」の保持が重要だ、ということである。この声明が起草された直接のきっかけは、アマゾンが人事採用に際して補助的に利用していた機械学習システムが、女性に対して不利益に働くことに気づき、その利用を停止したことに始まる。この事実が示唆しているように、大規模言語モデル（LLM）はインターネット上にアップされた膨大な言語データ（GPT-3は約4000億語に相当する文章を「学習」していると言われる）を収集・解析しているため、その中には社会にまん延している偏見や差別も当然含まれているし、場合によってはそれらを増幅させ拡散させているかもしれない。もちろん、明示的な名誉棄損やプライバシー侵害、あるいは個人情報流出や著作権侵害に当たる表現があれば、現行法でとりしまることができるが、既存のデータに紛れ込んでいる無意識の偏見やジェンダーバイアスについては、指摘されるまでは気づかないことが多い。

最近の新聞報道によれば、無料版のGPT-3.5を用いた調査で男女の職業親和性（看護師は女性向きなど）について問いかけたところ、バイアスが含まれる回答が4割に上ったという[10]。やっかいなのは、そのバイアスが「学習」され、強化された形で再生産されることである。しかも、それが時代の社会通念を反映しているとすれば、容易なことでは取り除くことができない。またこのような場面では、情報リテラシーや情報倫理教育が役に立つとも思えない。偏見や差別表現は、気づかれないままに膨大な過去のデータの中に埋め込まれているからである。それゆえAIの「再教育」が必要とされるゆえんだが、逆に生成AIを使って無自覚に広がっている差別や偏見を焙り出すことはできるかもしれない。そのためにもブラックボックスと呼ばれているAIの作動プロセスを透明化し、使用されている学習データの開示を進めるべきであろう。ただしこれは国際的な合意のもと

に進められなければ意味がない。抜け道を見つけては規制を強化するイタチごっこになりかねないからである。

4 生成AIは責任を取れるか

人工知能が実現する未来については「スーパーインテリジェンス（超知性）」「シンギュラリティ（技術的特異点）」「知能爆発」「ホモ・デウス」「トランスヒューマニズム」など百花斉放というべく、SF的議論が喧しい。だが、汎用AIができたとして、それは誤作動や事故に際して、人間に代ってさまざまな「責任」を取れるのであろうか。国際技術哲学会の会長を務めるマーク・クーケルバークは、アリストテレスの『ニコマコス倫理学』を踏まえながら、以下のように否定的議論を展開している。

> 機械は意識、自由意志、感情、意図を形成する能力などを欠いているのだから、機械は行為者ではありえても、道徳的行為者ではありえないのである。たとえばアリストテレスの見解では、人間だけが自発的な行動を遂行することができ、自分の行動について熟考することができる。もしそうだとすれば、唯一の解決法は、機械が行ったことに対しては人間に責任を取らせることである。人間は機械に行為者性を委託するが、責任は持ち続けるのである。[11]

たしかにこれは妥当な考え方であろう。先の西垣通もまた、「AIエージェント」という擬似人格の想定に対しては「行動の結果について倫理的／道徳的な責任を負えるのはあくまで人間以外ではない。つまり個人か、その集まりである法人に限られる」[12]と同様の考えを述べている。少なくとも「スーパーインテリジェンス」のような想定はSFの世界では許されても、現実世界においては妄想のたぐいにすぎないのである。

先の引用にあったアリストテレスの見解とは、行為者に道徳的責任を問えるのは、行為者の内にその行為の始まり（アルケー）、つまり意思決定の能力があり、しかも自分がなした行為について知っている場合に限られる、というものである[12]。たとえばAIを搭載した自動運転車が人身事故を起こした場合には、AIを逮捕・起訴できない以上、その責任は自動運転車のプログラムの作成者か、それを製作した会社か、あるいは同乗して車を制御できる立場にあった運転者か、いずれにせよ人間が取らざるをえない。

すでに現行の法律として成立している「製造物責任法（PL法）」の考え方がAIの場合にも参考になるかもしれない。クーケルバークによれば、「製造物責任法では、個人の過失は問題にせず、技術を提供した

企業が、その企業の過失の有無に関わらず、損害に対する賠償をしなければならない」[13]のである。しかしながら、AIに「法的人格（法人）」を認めることに対しては欧州議会（European Commission）を中心に根強い反対がある。責任の所在を不明確化するために濫用されるおそれがあるからである。AIは極めて有用な道具であるが、それを使いこなすためには、「AI＋人間系」の社会的整備のために人間も機械も知恵をしぼらねばならない。

おわりに

パスカルの『パンセ』（ブランシュヴィック版）の冒頭には「幾何学的精神と繊細の精神との違い」をめぐる断章が置かれている。

> 前者［幾何学の精神］においては、原理は手でさわれるように明らかであるが、しかし通常の使用からは離れている。したがって、そのほうへはあたまを向けにくい。慣れていないからである。しかし少しでもそのほうへあたまを向ければ、原理はくまなく見える。[14]

明らかなように「幾何学の精神」とは論理的推論能力のことであり、AIが最も得意とする領域である。正確さにおいてもスピードにおいても、とても人間の及ぶところではない。それに対して「繊細の精神」については、「このほうの原理はほとんど目に見えない。それらは、見えるというよりはむしろ感じられるものである。それらを自分で感じない人々に感じさせるには、際限のない苦労がいる」[15]と記されている。こちらの方はAIには委譲できない「人間にしかできないこと」を表していると考えられる。一言でいえば「共感力」と名づけることができるであろう。その点については教育工学者の美馬のゆりが以下のような的確な指摘をしているので参考になる。

> 後者［人間にしかできないこと］において私が注目するのは、近未来においてもAIの実現が遠いと考えられる、人間的な側面である「共感」です。共感は英語に訳すと”sympathy” あるいは、”empathy” と二つあることに気づきます。前者は日本語では「同情」と言い換えられます。後者には他者の立場に身を置き、積極的に相手を理解しようとする、といった意味が含まれます。つまり「共感」とは、他者の感情だけでなく、意図を理解し、共有することなのです。[16]

パスカルの言う「繊細の精神」を「共感力」と言い換えてみれば、「幾何学的精神」との対比がはっきりしてくるであろう。パ

スカルはそれに続けて「幾何学者が繊細で、繊細な人が幾何学者であるのは珍しい。なぜなら、幾何学者はそれらの繊細な事物までも幾何学的に取り扱おうとするからである」[17]と述べている。これなどはデジタル還元主義者への根本的な批判とも読むことができる。

生成AIと人間社会の付き合いは始まったばかりであり、共存の模索も現在進行中である。使い方さえ誤らなければ有用この上ない便利な道具だが、他方ではハルシネーション、偽情報、フェイク画像など民主主義の根幹を揺るがしかねない危険な要素をも潜在させている。パスカルではないが、AIの作動原理である「幾何学の精神」は、共感を基盤とする「繊細の精神」と相互補完的に協働し合ってこそ道具として十全な力を発揮することができる。AIの誤用や悪用を根絶することは困難であろうが、それはAIが「人間の鏡」[18]であるからにほかならない。AIとの共生がユートピアになるかディストピアになるかは、この鏡をどのように磨き上げるかにかかっているのである。

注

1　最新の新聞報道によれば（『朝日新聞』2023年12月1日付朝刊）、G7広島サミットで打ち出された「広島AIプロセス」の最終合意案がまとまったとのことである。まだ具体的内容については公表されていないが、どのように実効性が担保されるのか、今後の推移を見守りたい。
2　木田元ほか（編）『コンサイス20世紀思想事典 第2版』三省堂、1997年、「チューリング・テスト（野家啓一）」の項目。
3　土屋俊ほか（編）『AI事典』UPU、1988年、「フレーム問題（平賀譲）」の項目。以下の例も同項目による。
4　ジョン・サール『心・脳・科学』土屋俊訳、岩波書店、1993年、34-35頁。
5　同前、36頁。
6　同前。
7　『Newton別冊』「ChatGPT徹底解説」、ニュートンプレス、2023年10月、22頁。
8　西垣通・河島茂生『AI倫理』中公新書ラクレ、2019年、87頁。
9　人工知能学会のHPによる。
10　『朝日新聞』2023年11月21日付朝刊による。
11　マーク・クーケルバーク『AIの倫理学』直江清隆ほか訳、丸善出版、2020年、93頁。
12　西垣通・河島茂生、前掲書、141頁。
13　マーク・クーケルバーク、前掲書、131頁。
14　ブレーズ・パスカル『パンセ』前田陽一・由木康訳、中公文庫、1973年、7頁。
15　同前、8頁。
16　美馬のゆり『AIの時代を生きる』岩波ジュニア新書、2021年、74-75頁。
17　ブレーズ・パスカル、前掲書、9頁。
18　日本学術会議哲学委員会の主催で開かれた公開シンポジウム「AI時代における哲学・美学・倫理学・宗教学」（2023年11月25日、オンライン）での吉岡洋前委員長の閉会挨拶の言葉より。これに限らず、このシンポジウムからは数々の刺激と示唆を受けた。パネリストならびに企画・運営に携わられた方々にこの場を借りて御礼申し上げたい。

PROFILE

野家啓一（のえ・けいいち）

- 東北大学名誉教授
- 日本学術会議連携会員

専門
哲学・科学基礎論