武蔵高等学校中学校紀要 8：149–156（2024）

生成 AI、ないし大規模言語モデルの進化によって わたしたちが失うもの

水上　雄亮
国語科

われわれ人間は言葉を理解し、そして言葉を用いることができる。「言葉を理解できる」とは「言葉を用いることができる」ことにほぼひとしく、「言葉を用いることができる」は「言葉を理解することができる」ことにほぼひとしい。おそらく、これが我々の自然な言語観である。

しかし、この自明とも思える前提はもはや崩れ去ってしまった。あるいは、崩れ去るのも時間の問題だと思われる。そして、われわれが「言葉を理解できる」と主張することも、もはや許されなくなるだろう。

近年では、大規模言語モデルの進歩によって、AI がかなりの程度において人間の言語運用を模倣することができるようになった。もちろん、その度合いはいまだ完全とはいえない。しかし、われわれがその使用開発を制限したり、あるいは無制限な開発を憂慮したりするほどには、すでにその優位は揺らぎつつあるように見える。じっさい、積極的な制限や介入が行われない限り、人工知能が人間の言語運用と見分けが付かない、少なくとも人間の印象評価では判別ができないレベルの能力を獲得する日も、そう遠くないと思われる。すでに、将棋のような分野では AI が人間の能力を凌駕している。自然言語の運用についても、そうならないと考えることの方がよほど難しい。

ところで、いま「判別ができない」といったが、「人間の言語運用と見分けが付」くか否かについての客観的な判断基準は今のところ存在しない。というより、それ以前に我々は統語の一般的規則すら見出していない。つまり、どのような文が有意味であるか否かを決定しうるような明確な基準を、われわれはまだ知らないのである。どのような文が有意味かを正しく定義できていないのに、その文が正しい運用であると、どうして厳密に判定できるだろうか。

我々に可能なのは、発話された個別の文に対して、それが置かれたコンテクスト上において、妥当かそうでないかを判断することぐらいである。だから、「AI が人間の言語運用を

模倣しうるか」という上記の問いかけには、おそらく現状では古典的な「チューリングテスト」によって判断するぐらいしか方法はない。

「テスト」とはいうが、その内実はインターフェース上で「相手」と一定の対話を行い、その際の印象に基づいて、相手が人間であるか否かを判断するというものである。完全に主観的な印象評価にもとづく方法なのだが、これ以上の有効性を期待できそうな判定方法は、今のところ存在しないように思われる。「AI そのものにその判定方法を学ばせる」という手法は現に模索されているだろうし、実際その精度は人間のそれをすでに上回っているのだろうが、それは人間自身が判断することを放棄する、ないしは他の存在にゆだねる、という結論をあらかじめ前提にしている。

チューリングテストの「合格条件」を具体的にどう設定するかについては、様々な議論が可能だと思われる。たとえば、「その言語を母語とする話者のうちからランダムに１００人を無作為に抽出し、そのうちの７０％以上が認めれば、その被験主体は『言葉を理解できる』、『言葉を用いることができる』と判定してよい」といった具合であろう。何人から抽出するのか、その抽出条件は完全な無作為なのか、何％のラインを合格と見なすか、など具体的に詰めるべき諸条件はあるが、おおむねこのような方針で決定されるはずである。

ただし、これは倫理的な問題を生じさせかねない。上記のような条件を設定した場合、生物学的にヒトであることが確かであるにもかかわらず、チューリングテストを突破できない者も出てくるだろう。その場合、チューリングテストの実施そのものが新たな差別や分断を生み出しかねないという懸念がある。だから、現実的にはこのテストを実施すること自体が、困難かもしれない。

とりあえず、ここではテストを実施し得たという仮定のもと、議論を進める。おそらく、将来のどこかそう遠くない時点でチューリングテストに合格する人工知能が登場するはずである。対話が行われる状況を細かく限定すれば、突破はより容易になるだろう。対話時間を短くしたり、話題や文脈をあらかじめ制限したりすれば、自然な会話を成立させやすいからである。

そしてひとたびテストを突破してしまったならば、その人工知能は「言葉を運用することができる」とみなさなければならない。先ほども触れたが、言葉には「厳密に正しい運用」というものがない。我々の目から見て、一連の対話に違和感がないならば、その相手は言語を運用できているのである。

むしろ、さまざまな用例から考えるならば、「正しい運用」の基準というものはかなり柔軟に変化するし、かつては「不適切」とみなされていた運用の仕方が、後の時代には「正しい」範疇に収まったりもする。かりに統語の一般規則というものが見出されたとしても、少なくともその一部には自由変数的な部分があり、ルールの一部はたえず可変的だと考え

るべきなのだろう。

人間が少々破格な語用をしてもそれは認めるが、AI の場合は認めない、というのはダブルスタンダードである。そもそも、そのような恣意性を排除するために、テストはインターフェースを介して行われるはずだ。表示される言語運用以外に、相手のことを判断できるような情報は全て遮断される。その意味で、AI が多少不自然な運用をしたとしても、それがただちに判定に影響を与えるかは分からない。　むしろ人間の、対面での会話では問題とみなされないような「破格的」な語用が、「AI のようだ」と逆に疑われ、テストを突破しなくなる可能性さえ考えられる。

むろん、テストの結果を人間と AI とでグルーピングしたとき、「不自然な運用をしてしまう確率」が両者の間で有意に異なる、という可能性はあるだろう。しかし、それはあくまでも集団に認められる傾向の差異に過ぎないのであって、個々のテストを判断するための判断材料にはなりえない。判断に影響しうるのは「人間も AI も不自然な運用をするが、その不自然さにはそれぞれ特徴があり、有意に弁別することが可能である」という場合のみである。しかしその場合とて、それが人間には判別できず、AI にのみ判定できるような特徴の差なのであれば、少なくともチューリングテストの上では意味をなさない。

ともかく、チューリングテストによって一定以上の水準を超えたものは、「言葉を用いることができる」とみなす以外にない。これは、どうあってもわれわれが受けいれざるを得ないことである。じつはその先に、「人間には違和感がないように見えるが、AI の判断によれは人間の発話主体ではない」と判定されるグループが存在しうるだろうが、現在の議論とは問題の性質が異なるのでこれ以上言及しない。

しかし、その上で我々は、こうも主張することができる。「その人工知能が言葉を用いることができることは認める。しかし、それが言葉を理解できるとまではみなせない」と。

なぜそのような主張が可能なのか。それは、現代の人工知能が「言葉を理解する」という目的を放棄した上に成り立っているからである。

かつて、人工知能の研究は「人間の言葉を理解できる AI」を生み出すことを主要な目標の一つとしていた。人工知能とは、文字通り「人工的に製作された、知能のある存在」である。そして知能があるということと、「言葉が理解できる」ことはほとんど同義である。少なくとも、われわれ人間にとってはそうだ。自身のあり方からして、「言葉を理解できない、あるいは理解しない知的存在」というものを、われわれはうまく想像できない。だから、人工知能研究がそれを志向したのははいたって自然な成り行きだった。

しかし、ある時期から人工知能研究は「言葉を理解できる」という大前提の目的を放棄

してしまった。それどころか、その基礎となる「人間の思考を模倣ないし再現する」という目的すら、放棄しているのである。「言葉を理解できることを目指さない」ならもう AI 開発研究そのものを放棄したようにも思えるが、そうではない。それとは異なる形で、「言葉を用いることができる」AI の開発が進んだのである。

端的にいえば、新しいタイプの AI は「言葉を理解する」ことの代わりに、「それらしい言葉をうまく並べる」ことを目指した。「それらしい言葉を並べる」だけなら、言葉を理解していなくともできる。新しいタイプの AI は、膨大な「用いられた言葉」の記録を学習する。いわゆる、「コーパス」から可能な限り膨大な用例を拾い集め、その全体的な傾向や特徴を掴むのである。現在では、紙媒体に限らず、インターネット上からも膨大な記録を集めることができる。

その上で、新しい AI は言葉と言葉の「つながり」を統計的に、つまり「確率論的」に整理する。この言葉には、これまでの用例の集積からいってこのような言葉が連なりやすい傾向にある、という「経験則」を膨大にかき集めることで、それを参考にして、人間が見ても違和感がないような「それらしい言葉」を並べることができるのである。その経験則の構築に費やされたサンプルの数は、ひとりの人間がその一生のうちに見聞きするだろう用例の量を、遥かに凌駕している。つまり、AI という存在は、かつて歴史上存在した人間の誰よりも、話された用例についての広い見聞をもっていることになる。

そして、これまでに話された用例を参考にするならば、「私は」という言葉の後には、「空き缶です」という言葉よりも「日本人です」という言葉が連続する可能性が高い。また、「私は」「日本人です」という言葉が連続していれば、その次には「よろしくお願いします」という言葉が続く可能性が、「絶対に許さない」という言葉よりも高いと判断できる。もちろん、そこに語の意味の理解はない。あくまでも、統計的に依拠した経験則の寄せ集めにすぎない。

拙い説明だが、原理的にはこのようなやり方で現行の AI は会話「らしきもの」を成立させている。言葉を理解しているわれわれ人間からするとずいぶん迂遠というか、むしろ言語の運用という観点からすれば見当外れで不適切なやり方にも思えるが、現に対話可能な AI はこうした原理によって機能しているという。AI の「能力」が優れているほど、また依拠する言語モデルのデザインが優れているほど、この用例の集め方と整理の仕方、またデータの取り扱い方が洗練されているのだろう。つまり、「それらしさ」がより自然なものに仕上がっているのである。近年では、その効率的な「学び方」そのものも AI 自身が自己学習していくタイプが登場している。

しかし、いくら「それらしい」言葉を並べて「言葉を用いることができる」ように見える AI も、しょせんは「それらしく見える」だけのことであって、実際に「言葉を理解できる」

とまでは言えない。せいぜい、「言葉を理解できる」ことを装う、模倣することができるという程度のものである。なので、いくらチューリングテストを突破し、「言葉を用いることができる」と判断された AI も、「言葉を理解している」とはいえない、と断言してもゆるされるだろう。

しかし、実はこのことはわれわれ自身の言語観にとっても重大な帰結をもたらす。言葉の理解抜きに言葉を運用できる、少なくともそのように見える存在が誕生したとき、冒頭で述べた「言葉を理解できる」とは「言葉を用いることができる」ことであり、「言葉を用いることができる」は「言葉を理解できる」であるというわれわれの自然な言語観が、突き崩されるのである。AI の登場によって、「言葉を用いることができる」としても、「言葉を理解できる」とはみなせない、という反例が示されるからである。

「それらしい」言葉を並べることができるのだから、「言葉を理解できる」とみなして構わない、とただちに認めるのは困難だろう。つまり、「言葉を用いることができる」ことを示すだけでは、「言葉を理解できる」ことを示したことにはなり得ないのである。

「言葉を用いることができる」ことを示すだけでは「言葉を理解できる」ことが示せないというのであれば、そのことは我々の言語運用にもまったく同様に適用されなければならないだろう。われわれは人間だから「言葉を用いることができる」ことをもって「言葉を理解できる」ことを示したことになり得る、というのは安易な特権的思考であり、退けられるべきである。それを許してしまえば、人間が「言葉を理解できる」のは何故かというと、それは人間だからであるという同語反復が許されることになってしまう。

つまり、われわれは「言葉を用いることができる」という事実を示す以外の方法で、「言葉を理解できる」ことを示さねばならない。もしそうでないなら、「言葉を用いることができる」のであれば、用例と統計に基づく「ことばを理解できるかのように振る舞う」AI も、「言葉を理解できる」と結論しなければならないが、それはわれわれの自然な言語観に反している。「言葉を用いることができる」ことをもって「言葉を理解できる」ことが示せないのだとしたら、われわれも「言葉を用いることができる」ことを示す以外の方法で「言葉を理解できる」ことを示さねばならない。

このとき、われわれはつい次のように考えたくなる。「われわれが言葉の意味を理解しているのは経験的にいって自明のことなのだから、それを何らかの形で示せればよい」と。現に、この文章を執筆しているわたし自身も、執筆している内容を理解しつつ、文章を書

いている。少なくとも、理解していると自分では思いつつ、あるいはそのことを何ら疑うことなく、書いているのである。加えていえば、この文章を読んでいる相手もそのことを認めてくれるだろうと期待しつつ、文章を書いている。

しかし、「言葉の意味を理解している」ことを「言葉を用いることができる」こと以外の方法によって示すことは、じつは極めて困難である。なぜなら、「言葉の意味を理解している」というのは、突き詰めると本人のみに固有の感覚的事象、いいかえれば「クオリア」にすぎないからである。「意味を理解している」というこの感じ、自分自身に内的に生じた「理解できている」という感覚を、他人に伝達することは非常に困難、というより原理的に不可能である。それは主観的な、あるいは独我論的な感覚、ないし経験に他ならない。この不可能性については、すでに哲学史上において議論が尽くされているのでここでは繰り返さない。「言葉を理解することができる」という主張を人間がお互いに認めあうとき、実は両者はお互いに交換不可能なものについて語り合っているのである。いわゆる、独我論的なパラドックスが生じているといえる。

現代ではｆMRIなどの装置によって、脳のおけるある機能や活動の状態がこのような「理解」に対応している、という特定の「対応関係」が見つかる可能性もあるだろう。その場合、脳機能にこのようなパターンが生じていれば、その脳の持ち主は「言葉を理解してる」とみなすことができるかもしれない。このような特定の状態が、言葉の理解を生じさせている、あるいは同時的に発現しているという共起的関係である。

だが、これも次のような想定が可能である。AI にも、その対応関係をそっくり「学習」させることができれば、同じ「対応関係」を模倣的に生み出すことができる。つまり、AI がある言葉を用いているときに、別のインターフェース上に同等の脳機能の「パターン」が再現されるよう、学習させるのである。この場合、AI においても、見かけ上ある「言葉の用い方」とある「脳機能の状態」の対応関係が生じていることになる。

もちろん、AI は「脳をもつ生物」ではないので、その脳機能は「見かけ」に過ぎない。しかし、この「見かけ上模倣できている」という点はかなり厄介である。少なくとも、インターフェースを介して行われるチューリングテストにおいては、対応関係が見せかけかそうでないかを判断することはできない。

そもそも、新しいAIの「言葉を用いることができる」ことも見かけ上の模倣だったことを忘れてはなるまい。「見かけ」が偽りであり問題とされるのは、そこにあるべき「本質」が存在しないからである。しかし、その本質を容易には提示しえないことこそが、この議論における問題なのである。われわれは、自分自身が言葉を理解しているという体験を今まさにしていることを、それ自体として提示できないからこそ、苦しい立場に陥っているのである。

模倣は、ある「閾値」を超えると模倣される対象それ自体と見分けがつかなくなる。それらを弁別する一切の手段が失われたとき、「模倣」という言葉はその意義を喪失し、その後には判別不能な二つの「それ」がただそこにあるのみである。両者の根源や由来について、一つは初めからそれであり続け、またもう一つはある時点からそう成り代わったのだとしても、そのような過去は何ら意味をなさなくなる。

AI が明らかに「計算機」様の外見をしていて、われわれ人間と姿形が明白に違うのであれば、まだ良い。これは現状ではまだ可能性にすぎないが、もし「言葉を用いることができる」ように見える AI が、「外見だけでは人間と区別が付かない精巧なアンドロイド」の機能として実装され、なおかつそのアンドロイドの「脳波」を外から f MRI で計測すると、人間の脳を計測したのと同じようなパターンが表示されるとしたらどうか。チューリングテストを AI が突破したときと全く同型の問題が、さらなる難問となって、再び回帰することにはならないだろうか。

再び結論しよう。AI がチューリングテストを突破した場合、われわれ人間は自分たちが「言語を用いることができる」と主張することはできても、「言葉を理解できる」とは主張できなくなるだろう。つまり、われわれが「言葉を理解できる」というのは客観的事実として提示可能なことがらではなく、単なるそのような権利の主張であるか、あるいは根拠のない思い込みにすぎない。

なぜなら、われわれは「言語を理解できている」というクオリアを、他者と共有可能な形で外示することができないからだ。「言葉を用いる」ことが可能であり、かつ「言葉を理解できるとはいいがたい」ような存在が登場してしまったならば、「言葉を用いることができる」ことは「言葉の理解」を証明しないからである。

もちろん、われわれにはその根拠のない主張をお互いに認め合う、という選択も許されている。人間という存在を特権化し、人間は「言葉を理解できる」知能があるが、AI にはない、と恣意的に区別するような選択である。その場合、「知能がある」ことの条件には、「言葉を理解できる」ことの他に、「脳を初めとした身体をもつこと」という新たな条件が加わることになるだろうか。あるいは、「我々が「知性をもつ」とみなしうるような、そのように我々が共感しうる身体や振る舞いの特徴を持つこと」という条件だろうか。ただし、知能を持つこと以外の特徴を「知能がある」ことの条件にするのは、奇妙といえば奇妙ではある。

そしてこれは、AI の進歩ではなくアンドロイドの進歩に問題解決の場を先送りしたにすぎないだろう。将来、人間にごく近似した容貌を備えた AI が誕生したとき、我々はいよい

よ「逃げ場」を失う。見た目も人間とそう区別しがたく、その振る舞いも人間と区別しがたい、しかし人間でないことがその由来からして明らかな存在が登場したとき、われわれはわれわれ自身をどのように再定義するのか。そのとき、再びわれわれは「魂」の存在を持ち出すのだろうか。しかし、魂の実在をどのように示すというのか。

あるいは、ひるがえって「チューリングテストを突破しているなら、その AI は『言葉を理解できる』と結論しうる」と認めてしまうなら、話は全く別である。ただその場合、「言葉を理解できる」ことを模倣しているだけのはずだった AI に、「知能がある」ことも認めなければならなくなる。しかし、その瞬間からわれわれは AI という存在をどのように扱ったらよいのか。曲がりなりにも知能を備えた存在を、単なる機械や装置、つまり動産や器物の類として留めおくことは許されるのだろうか。知能があるという特徴は、たとえば情動をもつ存在である他の動物たちと比べて、どう位置づけられるべきなのだろうか。

「知性と、情動とを兼ね備える存在こそが人間なのだ」という新たな主張を掲げて、人間の特権性を再獲得すべきだろうか。しかし、知性を備えるに至った AI にとって、もう一方の情動を得ることはさほど難しくないと思われる。かんたんに言ってしまえば、ある種の頑迷さや偏差といった非合理的選好性があるだけでも、そこに情動があるとみなすには十分だろうからである。

われわれは、SF 作家や哲学者が純粋思弁の中でのみ格闘してきたような問題に、現実として向き合わねばならない時代を生き始めている。

付記

どうやら現実には、AI が生成した発話か否かを判定するのは、専ら AI 自身の「仕事」になるようである。とすれば、事態はすでにここで想定されているよりも、さらに憂慮すべき事態に至っているのではあるまいか。われわれは、本稿によって憂慮される「喪失」の経験可能性すら、すでに失っているのではあるまいか。AI の判断を尊重することが人間においては最も知的な態度となり得るような世界が、すぐそこまで迫っている。