技術情報

AIが人類の脅威になるという幻想

株式会社ウサギィ　取締役　五木田　和也

1．はじめに

昨今は「人工知能（AI）」が一大ブームとなっている。IBMのWatsonがクイズ番組で人間のチャンピオンに勝利したり，英DeepMindによる囲碁AI，AlphaGoが人間のトッププロに勝利したことは記憶に新しい。

AIという言葉は実に幅広い意味で使われている。音声認識もAI，画像認識もAI，車の運転もAI，株価予測もAI，なんでもAIである。一方，サイエンス・フィクションで描かれるような“いわゆるAI”もある。わかりやすい例でいえば，ドラえもんやターミネーターであろう。後々詳しく述べるが，前者のAIはご存知AI，後者のAIはAGIと呼ばれる全く異なるものである。

ここで扱いたいのは後者のAIであり，AlphaGoの強さの秘密や，車の自動運転技術最前線について語るつもりはまったくない。砕けた言い方をすると，「ドラえもん（のような知能）はいかにして作られるか？」がメインテーマである。

AI技術が進歩していくとターミネーターやスカイネットのようなものができ人類の脅威になるというのは本当だろうか？

2．AIとはなにか？

まずは「人工知能（AI）」と「機械学習」の違いについてだが，厳密な定義はさておき，AIは文字通りの人工的な知能のことである。知能というのは非常にややこしいもので，未だに厳密な定義がない。よく用いられるのは「未来を予測する力」や「環境への柔軟な適応能力」などである。大雑把に言えば，人が見て「知能がありそうだ」と思えばAIと言ってしまって構わない程度である。興味がある方は「中国語の部屋」という思考実験について調べてみていただきたい。

3．機械学習とはなにか？

現在流行っているAIは厳密にはこの機械学習の別名だと思って差し支えない。こちらは先ほどのようにゆるい定義ではない。与えられたデータを分析して，そこに役に立つルール・法則を見つけ出す，というのが機械学習がやることである。たとえば文字認識であれば，色々な文字を観察（分析）して，「“A”という文字は，このような特徴を持つ確率が高い」というルールを自律的に学ばせるわけである。囲碁であれば過去の対局を分析し，「この局面であれば，次にこのような手を打つとこれくらい勝率が上がるだろう」というルールを学ばせるわけである。

株価や天気の予測をしたり，囲碁や将棋をしたり，自動車の運転をしたり，絵を描いたりと，機械学習はなんでもできるようにみえる。この仕組についてすこし触れておく。まず，一口に機械学習といっても，大雑把に分けるとだいたい

－ 教師あり学習
－ 教師なし学習

－ 強化学習

の３つにカテゴライズすることが多い。この３つを順に見ていこう。

⑴　教師あり学習

教師あり学習はいわば機械学習の「花形」であり，単に機械学習といった場合は教師あり学習を指すことが多い。画像認識や文字認識といったものが代表例である。

筆者が勤めている㈱ウサギィも機械学習技術を核とする会社であるが，特に得意とするところはこの教師あり学習である。

これらはまず「この文字は“A”です」「この文字は“B”です」…といった例をたくさん人間が用意する。アルファベットの手書き文字認識をしたい場合は，1 文字あたり 100 ～ 1000 パターンは人間が教えてあげなければならない。このようなデータは教師データと呼ばれる。

教師あり学習を用いると，「このような特徴があるのならばＡという文字である」といったパターンを見つけ出せる。すると学習時には使わなかった未知の文字を入力しても，「このような特徴があるのならＡという文字だろう」という出力を行うことができる。

⑵　教師なし学習

明示的に教師データが与えられない場合はどうだろうか。たとえば，大量の商品データがあるとして，それを上手く階層構造をもって分類したいといった場合である（クラスタリングと呼ばれる）。

この場合教師データは与えることができない。なぜならここでの教師データは「どの商品をどうやって分類するか」であるが，今やりたいことは「たくさんの商品を自動的に分類したい」ことであるから堂々巡りである。このような場合はなにかしらの基準（ここは人が決める）をもって，その基準で最も良い結果が得られるように試行錯誤させるといったアプローチを取る。これが教師なし学習である。

⑶　強化学習

強化学習は「何をしたら良いのかはわからないが，ゴールだけは決められている」ようなパターンの問題で力を発揮する。

たとえば将棋やチェスを考えてみよう。ここでの勝敗はかなり後の方の局面（つまり詰みになる直前）にならないとわからない。それまでの局面では明確に「これが絶対に正解である」とは言い難い。「良さそう」「悪そう」といった大局観も経験を積むとある程度読めてくるが，それでも決定的なものではない。初心者なら勝って（負けて）みて初めて「あのときの一手は良かった」「あのときの一手のせいで負けてしまった」といった経験が形成されていく。

強化学習もまったく同じである。ここでも将棋やチェスを例にとれば，たとえば「勝利したら＋１点，敗北したら－１点」といったゴールだけを設定しておく。ゴールの指定のみで，その途中が良いか悪いかは人間は指定しないのがポイントである。

最初，強化学習のプログラムはなにもわからないので適当に指せる手を指していくだけだが，それだと当然負けてしまう。負けたときに「こういう手を指していくと結果的に負けてしまう」ということはわかるので，同じことはもうしないように行動を変える。強化学習のプログラム同士で戦わせていれば，1/2 くらいの確率で勝つこともある。その場合は「こういう手を指していくと結果的には勝つ」ということがわかるので，そういった手はもっと指すように行動を変える。これを無数に繰り返していくと，途中の状態に対しても「今は形勢がやや不利だが，こういう手を指せば良くなるはずだ」といった判断がある程度できるようになっていく。

このような試行錯誤によって，価値観や行動計画自体をまるごと学習させようとするのが強化学習である。

4. 特化型 AI と汎用 AI

ここまでで，今流行のディープラーニングを始めとする機械学習技術による AI と，ドラえもんやターミネーターのような AI はなかなか大きな開きがあることにお気づきかと思う。

あるデータに対して「このような傾向がある」とか「このようなデータに似ている」ということがわかるのはそれはそれで有用だが，それができたからといってドラえもんのような AI が作れるかというとかなり難しそうに思える。とてつもなく囲碁が強い AI ができたからといって，我々と同じような知能を持っているかというとそうは思えないだろう。砕けた言い方をすれば，そのような要素技術をもって「心」や「意識」が宿るかと言われると首を傾げざるを得ない。

AI 研究は当初は「人間のような知能をもった機械を作りたい」という率直でわかりやすいところから出発しているが，想像以上に困難であることが判明した。そこで，「人間ができるある特定の問題」だけに集中的に取り組み，実用的なものを作るという方向に舵を切ったわけである。たとえば，文字認識，音声認識，顔認識，チェスや囲碁をプレイする…といったところである。これらは言うまでもなく成功を収めている。

一方で，いつの間にか「人間のような機械」のほうはもはや忘れ去られてしまった。まず心や意識とは何か？といったところからほとんど解明されていないし，脳の全貌もまだ人類は把握できていない。

しかしながら今でも「いわゆる AI ではない，人間のような知能を持った機械」について真面目に取り組んでいる潮流も存在する。こちらは「AI」ではなく「汎用 AI（汎用 AI，AGI）」と呼ぶ。より明確に，今主流の AI のほうは「特化型 AI（特化型 AI）」と呼ぶ。

つまり，今の AI というのは，囲碁とか，音声認識とか，顔認識とか，個々の問題に“特化”している AI である。一方で我々人間を含めた高等動物の知能というのは，ある特定の問題に特化しているわけではない。生まれた直後はなにもできないかもしれないが，逆に言えば学ぶことでなんでもできるわけである。チェスに関する知識は遺伝子というプログラムに組み込まれているわけではないが，生まれたあとに学ぶことで習得できる。一方で既存の AI，つまり特化型 AI は，プログラムされた以上のことはなし得ない。

勉強にしても運動にしてもなんでもそうであるが，我々は相当な汎用性を有しており，あとからなんでも学習することができる。たとえば Watson はクイズ番組で人間に勝利したが，あくまでもクイズに特化したシステムである。逆にネズミやカラスは数学やチェスやクイズはできないが，驚くべき高い知能・汎用性・適応力を持っている。数学やチェスのような難しい問題を解くプログラムは簡単に実現できるのに，未知の環境で動き回るような単純な知能を実現するのは極めて難しいのである。この「動物にとって簡単なものほどコンピュータでは難しく，動物にとって難しいものほどコンピュータでは簡単に解ける」という逆説はモラベックのパラドックスと呼ばれている。

動物のような汎用性は今の AI には存在しない。たとえどんなに賢い囲碁の AI であっても，将棋ができるようにはならないし，車の運転を教えることもできないのである。

これが「特化」と「汎用」と呼ばれる所以である。「心」や「意識」の問題はとりあえず横においておくにしても，少なくとも高い汎用性

は極めて重要である。

5. 汎用 AI とは何なのか

汎用 AI はまだ開発に成功しておらず，それどころかどういった方針で研究をすれば実現可能なのかも不明瞭である。しかしながら一部の研究者は粘り強く研究を進めており，機械学習と脳科学を融合していく形で汎用 AI の実現を探っている。私が所属している特定非営利活動法人全脳アーキテクチャ・イニシアティブ（WBAI）はまさにこの汎用 AI の実現に向けた活動をしている団体であり，国内の AI 研究者が数多く参加している。以降では汎用 AI 研究の中でも比較的妥当性が高く，それなりに支持されている考えをもとに「いかにして汎用 AI を作っていくか？（と研究者は考えているか）」を紹介していきたい。

6. 我々の脳はどうなっているのか

ある程度知性が高いと思われる動物が共通して持っている脳の組織や構造を洗い出していくと，特に以下の 3 つのモジュールが重要そうである，ということがわかっている。

- 大脳基底核
- 小脳
- 大脳新皮質

(1) 大脳基底核

大脳基底核は，簡単に言えば「自分にとって得なことはもっとしたい，損なことはもうやりたくない」を管理しているところである。たとえば「この場所にいくと餌が食べられる」「この場所に行くと天敵に襲われるかもしれない」といった損得に関する事柄は，生き物にとって極めて重要である。人がギャンブルにハマってしまうのも大脳基底核のせいである。

(2) 小脳

小脳は俗に言う「身体で覚える」というのを司る部分である。たとえば自転車に乗るというタスクは最初非常に難しく，全神経を集中しないと行えない。パソコンのマウス操作も，初心者のうちはどうにも思った通りにマウスポインタが動かせずいらいらする。

しかし慣れてくると実にスムーズに身体が動かせるようになるどころか，考えなくても勝手に身体が動くように感じられる。十分に習熟した人にとっては，自転車に乗りながら他のこともこなせる。また，普通の知識と違ってこのような行動は一度マスターすると絶対に忘れない。久しぶりに自転車に乗るからといって，乗り方を忘れるということはないし，水泳も夏になるたびに学び直さないといけなくなるわけではない。小脳はそのような処理をしてくれている。

(3) 大脳新皮質

大脳新皮質は実に幅広い役割を持ち，ヒトが他の動物と比べて特異的に大きな表面積をもつ部分である。役割としては五感の処理，つまり「見る」とか「聞く」とかいった部分であるとか，言葉の理解や発話，短期的な記憶や，逆に中長期的な計画を管理したりするといったいかにも「知能」という仕事をしている部分である。

7. 脳に機械学習を対応付けられるか

幸いにも，先に紹介した 3 つの脳の器官はそれぞれ対応する機械学習技術が提唱されている。

- 大脳基底核 ⇔ 強化学習
- 大脳新皮質 ⇔ 教師なし学習
- 小脳 ⇔ 教師あり学習

となる。概要としては，

- 大脳基底核（強化学習）は快不快を元にして，自分にとって最も利益があるような行動を選ぶ
- 大脳新皮質（教師なし学習）は外界から得た複雑に変化する情報を整理して，本質的な情報，高次の情報を抽出する
- 小脳（教師あり学習）は教師信号が得られるものを担当して，大脳でやっている処理を代わりに請け負う

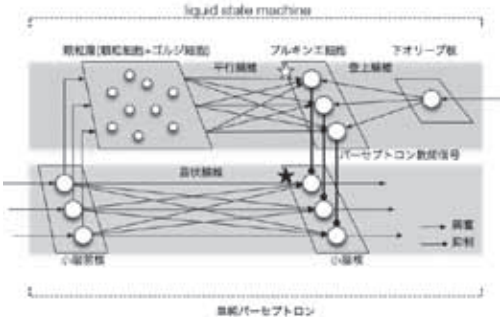
もちろんこれは非常に単純化されたものではあるが，一定の支持を得ているモデルである。

特に大脳基底核が強化学習であること，小脳

が教師あり学習であること，という仮説は神経解剖学的にも比較的妥当性が高い。

たとえば下の図は山崎匡らの「運動記憶の転送を行う小脳のスパイキングネットワークモデル」からの引用であるが，機械学習技術を使って小脳の構造の説明を試みている。

ここでは触れないが，記憶などを司る海馬についても様々なモデルが提案されているし，意識そのものを工学的に扱おうとする潮流すらある。下火であることは否定できないが，それでもなおゆっくりと歩を進めているのが汎用 AI の研究である。



8. さいごに

技術的特異点（シンギュラリティ）という言葉も最近よく聞くようになった。これは人間と同等以上の AI が生まれたときに，その AI がさらに洗練された AI を自ら設計し，そうして生まれた AI がさらに洗練された AI を設計し…というプロセスがひとたび始まると，AI は人の手を借りずともどんどん賢くなっていき，人間が取り残されるというようなシナリオのことを言う。ここでの AI は言うまでもなく汎用 AI のことを指す。自分自身の問題点を見つけ，設計・製造できるようなものは高い汎用性が必須となるからである。

このシンギュラリティが起きるのは，アメリカの科学者レイ・カーツワイルによると 2045 年ころになるという。つまりそれまでには汎用 AI ができるということである。遠い未来の話ではない。わずか 30 年もしない先の話である。

昨今の AI ブームに便乗して「もはやシンギュラリティは目の前であり，人類の危機となる」といった言説もみられる。既存の機械学習は大変有用であり，目を見張る速度で進歩しているが，この延長線上には汎用 AI のような「本物の」AI ができるわけではない。今の機械学習技術そのままでは汎用 AI には到達しえないのであるが，世の中にはこの 2 つを混同した言説が実に多い。

最近では，総務省が主導して行っていた AI 開発のガイドライン策定メンバーから Preferred Networks（PFN）が離脱してしまった。PFN はあまり知られている企業ではないかもしれないが，機械学習界隈では国内有数の技術力と知名度を誇る企業であり，海外においてもよく知られた存在である。そんな PFN が離脱した理由も，総務省の想定する AI が汎用 AI であり，全く話が噛み合わなかったことが一因であると言われている。つまり，国としては「画像認識や自動運転技術が発達しすぎてしまうと，ターミネーターのような恐ろしい AI が出来てしまう」…と考えていた（いる）わけである。本稿をご覧になった方であれば，まったく論点がずれた話であることは容易におわかりになるかと思う。PFN の言葉を借りれば「汎用 AI の実現は，今見えている技術の延長上には無い」ということだ。

一方で，汎用 AI について精力的な研究は細々と続いており，こちらも最近の AI ブームに乗って注目されつつある。AI 技術のブームによって，過剰に期待が高まってしまったり，人類の脅威になりうるといった意見が世間を賑わせているが，その認識は AI 研究の側から見るとかなりずれている。重要な技術であるがゆえに，今必要なのは正しい認識である。

本稿によってすこしでも AI，あるいは汎用 AI への興味を持っていただける方が増えれば幸いである。