Human-in-the-LoopのAI技術を目指して

Aiming at Human-in-the-Loop AI

国立研究開発法人産業技術総合研究所　フェロー　人工知能研究センター　研究センター長
辻井　潤一

国立研究開発法人産業技術総合研究所　フェロー、人工知能研究センター　研究センター長、英国マンチェスター大学教授、国際計算言語委員会（ICCL）委員長、AAMT／Japio 特許翻訳研究会委員長

1 はじめに

DX（デジタル・トランスフォーメーション）と脱炭素社会の構築は、近い将来の社会の在り方を考える上での２つの中心的な課題になっている。私がセンター長をしている産業技術総合研究所・人工知能研究センターでは、DX の中心技術の一つである人工知能（AI）技術の研究を推進している。また、産総研には、ゼロエミッション国際共同研究センターが設立され、ノーベル賞受賞者の吉野博士をセンター長として、脱炭素社会のための技術開発が行われている。いずれのセンターも、単なる産業振興だけでなく、社会全体のあり方を変え、人類社会への貢献を目指す中で新たな産業技術を生み出す研究を行っている。

本稿では、DX 時代の AI 技術が目指す方向、これからの科学技術の在り方について、私見を述べる。

2 AI と人間

人工知能研究センターの設立は 2015 年であり、この６年余りという短い間に AI 技術は大きな変貌をとげてきた。センター設立の当時は、人工知能の能力を人間と比較することで、その可能性と影響の大きさが議論されていた。

たとえば、Google 傘下に入った英国の DeepMind が開発した AlphaGo が、人間のプロ棋士に勝ったのが2015年の10月であった。翌年の３月には、韓国のトップ棋士、さらに 2017 年には中国の世界トップの棋士を破るということで、AlphaGo は、人間のプロ棋士では全く歯が立たない能力を示すようになった。韓国のトップ棋士との対戦が行われていた時、私も別の AI 国際会議でソウルに滞在しており、この対戦が引き起こした韓国社会の驚愕を現地で味わうことになった。

この 2015 年以降の AlphaGo の活躍は、2005 年のレイ・カーツウェルの「2045 年に人間を超える知性体の出現を予測する技術的特異点―シンギュラリティ」の議論と結びついて、「AI は人間を超えるのか」と、AI 研究者の中でも真剣に議論する人たちが現れることにもなった。

この種の熱に浮かされたような議論は、AI 技術が日常化し、社会に広がるにつれて変化してきている。むしろ、AI が社会の幅広い局面で使われることにより引き起こされる労働市場の変化、個人生活への AI の関与による監視社会の出現への危惧、AI がもつバイアスなど、より具体的な課題が議論されるようになっている。ただ、AI の判断を絶対的なものとしてみる AI 信仰や、逆に、AI を人間疎外の技術とみて排除する AI 警戒論の二極化は今も根強く残っている。よりバランスの取れた見方からの議論が必要になっている。

3 人間を超える AI の出現

大規模なデータと機械学習に基づく現在の AI 技術に道を拓いたのは、その前に起こったビッグデータ解析のブームであった。このビッグデータ解析の時代には、サイバー空間に集積されているデータを解析したり可視化

したりする技術、あるいは、そのデータの中に潜む規則性を取り出すデータマイニングの技術が研究された。

ビッグデータ解析では、可視化やデータマイニング技術がその典型であるように、膨大なデータに潜む傾向や規則を取り出し、それを人間が解釈し、それに基づいて人間が判断や戦略を立てることに焦点があった。

膨大なデータには、そのデータを作り出す背後にある種の機構が存在するはずで、それを人間が把握するのを助ける技術が、ビッグデータ解析であった。人間は、この技術を使うことで、自らの判断や戦略決定をより合理的におこなうことができる。エビデンスに基づくことで、合理的な判断や戦略決定を行える。

これに対して、現在の AI を特徴づける機械学習の技術は、データに潜む規則性を膨大なデータから把握し、それを内化し、判断に結びつけるところまでを計算機が行う技術である。AlphaGo は、過去の棋譜から学習し、その結果を使って、碁を打つプログラムを作り、さらには、このプログラム同士が計算機の中で模擬的なゲームを行うことで、膨大な量の棋譜を作り出す。このような人間がそのままでは解釈できない膨大な棋譜データを使って学習することで、人間の棋士の判断能力を凌駕する能力をもった。

膨大なデータを整理して人間に示すことで、人間の判断をより合理的なものにすることから一歩進んで、膨大なデータから学習し、判断までを行う AI プログラムが出現することになった。膨大な学習データから学習することで、自ら判断し碁を指す AI、戦略選択までを行う自律的（Autonomous）な AI ができた。

判断に人間の介在を不要にして、自らが判断する AI の出現である。碁を指すという能力だけを見ると、人間の名人よりも優れた判断能力と戦略選択能力を有する AI が出現することになった。

単純化すると、ビッグデータ解析では、人間が判断の主体であった。これに対して、機械学習に基づく AI では、判断の主体が計算機プログラムに移行した。これが、AI とビッグデータ解析技術の大きな差である。

4 AI の自律性

エビデンスに基づく判断は、エビデンスを無視した、思い込みに基づく判断や戦略よりも、より合理的なものであることは、一般論としては正しい。ただ、この論理を飛躍させて、膨大なデータから学習する AI の判断は、膨大なデータというエビデンスに基づいているがゆえに、人間の判断よりも常に合理的なものとなる、というのは間違いであろう。

まず、第一に、碁では、盤面にすべての情報があり、過去の棋譜や模擬ゲームの棋譜としてデータ化されている。これに対して、現実のタスクでは判断に必要なデータが限定できず、判断に影響するものすべてがデータ化されていないことも多い。このことは、現在、コロナ禍の中で、同じデータを見ても専門家の予測や判断に大きな差がでることからも、明らかであろう。人流や感染者数など、膨大なデータが集積されても、データ化されていない未知の要因や、データの背後にあるメカニズムの変化（例えば、変異ウィルスの出現）がある。これは、碁のような規則が明確な世界から、現実の問題に向かう場合には、しばしば起こることである。

第 2 に、合理的判断かどうかは、評価の基準に依存する。碁の場合には、ゲームの勝ち負けという評価基準が共有されており、勝ちに至る確率の高い指し手を選択することが合理的な判断となる。これに対して、現実の問題では、この合理性の基盤となる価値が必ずしも明確ではないことが多い。

この 2 つ、AI が判断の拠り所とするデータに関する問題と、合理性の基盤となる価値の問題が、AI の社会実装にとって考えるべき課題となっている。前者のデータに関する問題には、学習に使われたデータに偏りがあるとか、すべての状況を被覆していないなど、多くの解決すべき問題がある。

少し、価値の具体的な例として、医療分野での治療戦略の選択を考えてみよう。合理的な戦略選択の基準として、平均の生存期間が長いことを基準すると、外科手術の治療方針を選ぶことになる、とする。ただ、この外科的な治療方針では、直後に死亡する率もある程度あるとなると、平均生存期間は短くても、より安全な内科的治療を選ぶ人もいるであろう。それ以外にも、生存期間中の生活の質（Quality of Life）やその経済的コストなども、判断に影響する大きな要因であり、どの要因を重視するかで、合理的な判断は個々のケースに依存する。

この例のように、合理的な判断の基盤となる評価の基準が、判断によって影響を受ける側（例えば、患者やそ

の近親者）の価値観に左右される場合には、一律な判断をAIが行う自律性は問題になる。また、上のコロナ禍への対処のように、最適な戦略の選択に、データ化されていない要因が関与することも多く、観測されデータ化されたものだけで判断を行う閉じた自律系では対処できない問題も多い。

5 AIのブラックボックス性と人間との共同

AIの自律性が問題となる大きな理由に、AIのブラックボックス性がある。機械学習の中でも、現在のAI技術の中核となっているニューラルネットワーク、深層学習にこのブラックボックス性が強い。

ブラックボックス性とは、AIが下した判断や戦略選択の根拠が外部の人間からは不明である、ことをいう。

膨大なデータに潜む、人間の判断にとって有用な規則性を提示していくことで、人間の解釈や判断をより合理的なものにしていこうとしたのがビッグデータ解析であった。膨大なデータというブラックボックス性の高いものを、人間にとって解釈しやすい、有用な形に整理するというデータマイニングやデータ可視化から、機械学習の技術が発展する中で、膨大なデータをもとにした判断へと発展してきた。そのような過程でも、判断の過程を理解しやすい形で内的な構造化を行う決定木の手法などが開発されてきた。

ただ、判断の過程を人間にわかりやすい形で内化することと、判断の精度を上げることとは、相反する方向性を持っている。データが持つ、判断に有効な規則性を最大限に活用して精度を上げる方向に向かうと、ブラックボックス性が強くなる。

この典型が、ニューラルネットワーク、深層学習の手法になっている。深層学習では、データが持つ判断に有効な規則性が、モデル中の数百万－数億個のパラメータ（単純にいうと、モデル中のニューロン間の結合の重み）の形で内化されている。したがって、AIがある判断を行った根拠が人間にとって理解できない、ということになる。膨大なデータが持つブラックボックス性が、学習結果のモデル中の膨大なパラメータ集合に移行し、外部の人間には、判断結果だけが示されることになる。

これが、前節の自律性の抱える課題（データと価値）と組み合わさって、現在のAIの応用分野を狭いものにしている。たまたまデータに偏りがあったり、不完全な被覆があったりすることで、誤った判断を下している可能性があっても、それを検証することができない。対象を理解している人間からすると、ナンセンスなデータの偏りに判断が左右されている可能性もある。

よく挙げられる例に、動物が写っている写真とそうでない写真とを一見うまく識別できる深層学習モデルがある。このモデルは、実は背景がぼやけているかそうでないかを基準に写真を分類していた、という。実際、動物の写っている写真では、動物に焦点が当たっているために背景がぼやけていたので、そういう特徴が識別に役に立っていた、というわけである。これは、訓練データの偏りが判断に影響していた例で、訓練データに、車や人などに焦点が当たっていて背景がぼやけた写真が大量に入っていれば、このような誤りは起こらなかった。

ここでのポイントは、ブラックボックス性の高い手法では、このようなナンセンスな誤りが入り込んでいても、外からはわからないこと、ブラックボックスのAIの判断を絶対的な判断ととらえることの危険を示している。

また、ブラックボックスAIが、特定の価値観を合理性の基盤に置いていて、別の価値観からの判断では非合理的な判断をしている可能性も排除はできない。

複雑な判断や戦略の選択には、このようなブラックボックス性が常に付きまとう。人間の専門家の判断においても、外部の人間から見るとブラックボックス性は高い。ただ、AIと人間の専門家との大きな差は、人間の専門家の場合には、必要に応じて、自らの判断の根拠を（完全にではないが）説明することができる。また、説明の過程で、異なる価値観からの判断も提示することができることである。

複雑な判断の過程をできるだけホワイトボックス化することで、AIと人間との共同作業を促進させること、これが説明できるAI（Explainable AI - XAI）として現在、我々の研究センターでも、研究を進めている。

6 AIと責任

XAIの議論は、AIと責任の議論とも関連が深い。自律性を持ったAIが下した判断が誤っていて、事故や経済的な損失が生じた場合の責任をだれがとるのか、の問題は、多くの要因が絡み、今後、議論がさらに活発にな

る、と思われる。

当該の AI システムを開発した Vender、AI を訓練するデータ提供者、当該 AI システムを使ってビジネスを行っている人、当該 AI システムの使用者など、異なったステークホルダーが関与して AI システムが構築され使用されることから、その責任の所在を同定することは容易ではない。

学習する前のシステムの不具合、学習に使われたデータ集合の不備、特定のデータアイテムが引き起こした誤判断など、ある判断に関与した要因がトレースできることが、責任の所在を明確にするには不可欠となる。このトレース可能性（Traceability）は、XAI と類似の問題（AI のブラックボックス性）と深く関係している。

また、医療診断の例では、AI の判断結果を追認して医療行為を行う医師（AI システムの使用者）の責任も問題となろう。この場合には、AI システムが自らの判断根拠を医師にある程度説明し、医師自らがその説明に納得して AI の判断結果を採用できることが、不可欠となる。AI が下した判断を盲目的に採用する運用では、医師が責任をもって判断に関与することができない。XAI の技術が、不可欠となる。

医療と同じようなクリティカルな AI 応用に、人事評価や与信評価（Credit rating）の応用もある。この場合には、個人の利害に直接影響がある判断を AI が行うことになるために、不利益を被る評価を受けた個人に、その評価の根拠を開示する必要がある。データの誤りのために不利益を被る可能性をチェックできるなど、ここでも、判断過程のトレーサビリティと判断根拠の説明が必要となろう。

AI の自律性は、人間の労働コストを下げる AI の最大の利点である。実際、現在、自律性を持つ AI が社会に広がることで、人間が行うよりも誤りが少なく、また、コストも下がる場合も多い。自律性は AI を使う最大の利点でもある。AI 活用においては、自律性のある AI に任せられる AI 応用と、Human-in-the-Loop で AI のブラックボックス性を軽減することが必要な応用とを慎重に判断していく必要がある。

7 人間と AI の共進化

Human-in-the-Loop は、前節で議論したような AI システムの運用時だけではなく、AI システムの構築時においても重要となる。

AI は、摩訶不思議な魔法の杖ではない。AI システムを構築するためには、AI に行わせようとするタスクについての知識を、AI に与えていくことが必要であり、AI を人間が訓練する必要がある。

この AI システムの訓練に結構な時間とコストがかかる。AI は、導入した次の日から、コストが削減されたり、生産スピードが格段に向上したりする魔法の杖ではない。我々の研究センターは、あるタスク用の AI システムをできるだけ簡単に構築するための技術の体系を作ろうという、大型のプロジェクトを推進している。

AI システムに訓練する段階では、当該のタスクに関する知識を備えた専門家が Human-in-the-Loop の形で従事する必要がある。現在、この人間が持つ知識を AI 側に移行するためには、大量の訓練データを準備する必要がある。この訓練データの作成は、入力の観測データとそれに対する望ましい判断とを対応付けるアノテーション（付記）という作業になる。言い換えると、データとそれに対する正しい判断の対を訓練のために専門家が用意することが必要となる。このアノテーションのコストが大きくなり、AI の社会実装をする上での大きな障害となっている。

単純化していうと、専門家の役割は、データとそれに対する正しい判断の対をたくさん用意することで、その後は、AI の学習能力に任せる、という形をとっている。

訓練データの準備と AI システムによる学習の過程が切り離されている。この開発過程においても、AI と分野の専門家が緊密に協力する Human-in-the-Loop の考え方を積極的に取り入れようというのが、AI と人間の共進化という、我々の研究センターが推進しているプロジェクトである。

専門家が、観察データに対して正しい判断結果を与えた訓練データを作成する、と言っても、簡単なことではない。複雑な課題では、専門家の判断そのものが個人によって変わることも多い。例えば、画像からのがんの病理診断では、熟練した病理画像の専門家と、そうでない一般の医師の間では、30－40% の画像で判断の差が

出るといわれている。また、病理画像による診断の専門家の間においても、これほど、大きな差ではないが、かなりの差がでる。観察データに正しい判断をつけた訓練データの作成は、それほど、簡単な作業ではない。

病理画像からのがんの診断では、判断が分かれる微妙なケースでは、複数の専門家がコンファレンス形式で、各専門家が画像のどの個所に注目して判断したかを議論し、最終の判断に至る場合も多い。

このことは、専門家によって画像という観察データのどのような特徴に注目して判断しているかに差があることを示している。

病理画像を見て、その判断だけを付与するのではなく、どこのどのような特徴をみて判断したかを付記すると、さらに性能が上がるのではないか？　また、AI 側がどのような特徴をとらえて判断したかを人間に示すことができれば、AI の判断がナンセンスな特徴をみて間違った判断をしていないかを確認できるだけでなく、人間の専門家が気がついていなかった、判断の根拠となる特徴を見つけ出すこともできるのではないか？　Human-in-the-Loop で AI システムの構築過程に専門家との協力を密に行うことで、AI と人間（専門家）のいずれもがともに進化していくこと、が我々のセンターが推進している共進化 AI の研究プロジェクトである。

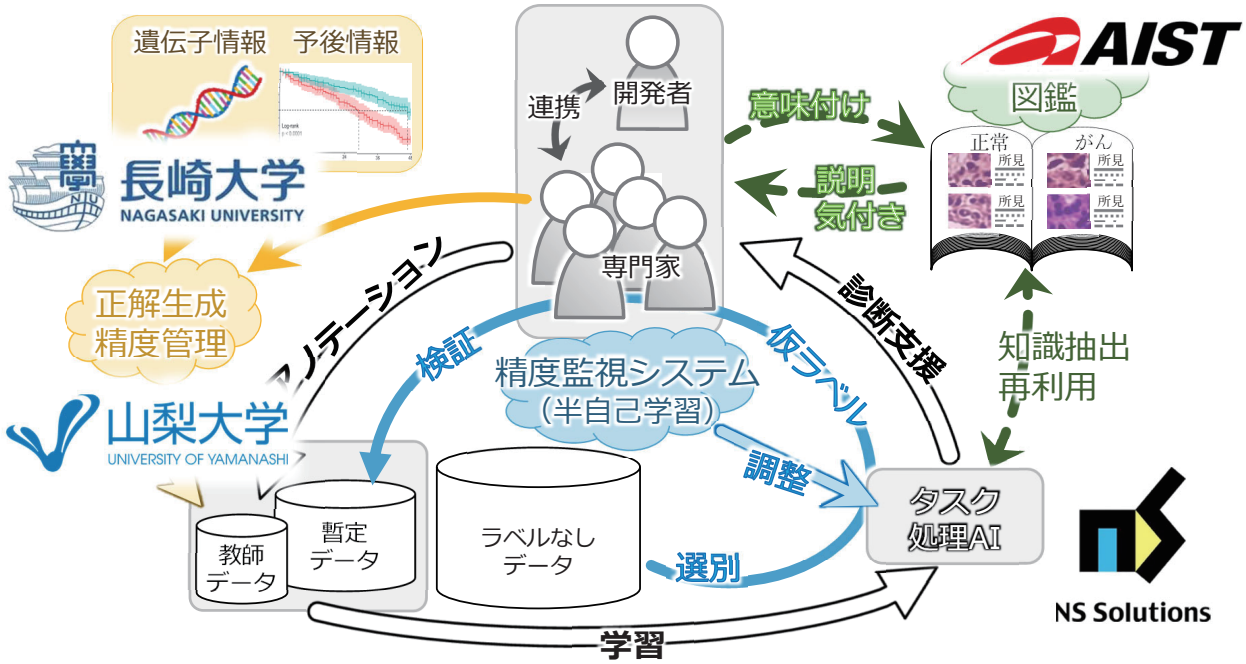
実際、このプロジェクトでは、病理画像の専門家と AI とがシステム構築時点で密に協力していくことで、大きな性能向上が得られている。また、病理画像と患者のほかの診断データとを結びつけることで、専門家が気の付いていなかった画像上の特徴で、AI が見つけ出した特徴がデータの偏りからのナンセンスな特徴なのか、そうではなくて、何らかの医学的な意味づけができる特徴なのかを検証していく研究にも着手している（図 1）。

8 おわりに

本稿では、膨大な量のデータを解析して、人間が解釈できる形に構造化して見せるビッグデータ解析の時代から、データに基づいてデータの持つ規則を学習し判断を行う機械学習の技術が発展し、その典型として、現在のニューラルモデル、深層学習があることを概観した。

データというエビデンスに基づく判断は、思い込みに

AIが獲得した情報（内部状態）を可視化しつつ蓄積し、専門家との合意形成を促す。整理された「判断根拠図鑑」を、判断根拠として提示。半自己学習によるモデルの訓練に再利用し更なる性能向上を図る。病理診断支援分野で医師との協働により実証。



【中間目標】各サブプロジェクトのプロトタイプを構築し、データセットの構築からタスク処理AIの訓練、判断根拠図鑑の作成、判定結果の説明までの過程で相互連携可能である事を確認する。本研究計画に参加する医師による性能評価を実施し、工学的および医学的な観点から個々の要素技術に関する課題抽出を実施する。

図 1　人と共進化する AI

基づく判断よりも、より合理的なものになるという主張は、一般的に正しいと考えられる。ただ、この議論をさらに拡張して、膨大なデータに基づくAIの判断は常に人間の判断よりも合理的である、とするのは間違いであることを、いくつかの例で議論した。データの偏り、合理性の基盤となる価値観の多様化など、AIの判断が必ずしも正しいものとは限らない。

もちろん、多くの人はAIの無誤謬性を信じているわけではない。ただ、依然として、データに基づく判断の重要性が主張され、その延長に、だからAIの判断は人間の判断よりも合理的という議論に向かうことがある。

理想的な形態は、人間とAIがお互いの知性の優れているところを取り入れていくことで、全体としてより合理的で、かつ、広い社会に受け入れられる判断、戦略選択に向かうことであろう。そのためには、AIの判断過程の説明可能性、理解容易性を向上させること、Human-in-the-Loopの考え方をより積極的に取り入れていくこと、であろう。

このような考え方は、気象変動に対する対処戦略の選定のように、多様なビッグデータの相互関連を読み解くという人間の専門家だけでは対処できない大規模な問題をビッグデータ解析とAI技術とを駆使することで、より合理的で、広く受け入れらる戦略を作る上では必須であろう。