「なぜ？」に答える説明可能AI

◆不正検出の理由を示すAI、死亡率と血中成分の関係をシンプルに示すAI

2020年5月、マイクロソフトは、同社のAIツールによって、航空会社のポイントプログラムを不正利用するユーザーを検出し、告発したことを明らかにした。この不正検出AIは、機械学習によって航空券の予約や支払いや特典の申請などの各種データを解析し、不審な行動パターンを見出している。不正を検出した際には、どの入力データの影響度が大きかったかを示し、AIが不正と判断した根拠を、人間が理解できるようにしている。

20年5月、中国の華中科技大学などの研究チームは、血液の成分データから新型コロナウィルス感染者の死亡率を予測する簡易な手法を開発した。研究チームは、武漢市の感染者の血液成分と死亡率のデータを、人が解釈しやすい予測手法を用いたAIで評価し、死亡率との関連性の高い血液成分を特定した。さらにこのAIを改良し、97%以上の精度で10日後の死亡率を予測する、3種類の血液成分と判定フローチャートの組み合わせを見出した。

AIの判断や予測の根拠を、人が理解できるように示す技術は、説明可能なAIや説明できるAI（両者合わせて、以下、説明可能AI）と呼ばれている。

◆AIのブラックボックス問題を解消する説明可能AI

AIの産業利用に向けては、AIがブラックボックス型のプログラムであることが課題のひとつとされている。第3次AIブームと呼ばれる昨今のAI利用においては、深層学習が最も注目され牽引役となっている技術であり、予測精度の高さから、画像認識や自然言語処理などの分野で広く活用されている。19年7月に特許庁が公開した「AI関連発明の出願状況調査報告書」によると、17年に出願されたAI関連発明のうち半数近くが深層学習に言及し

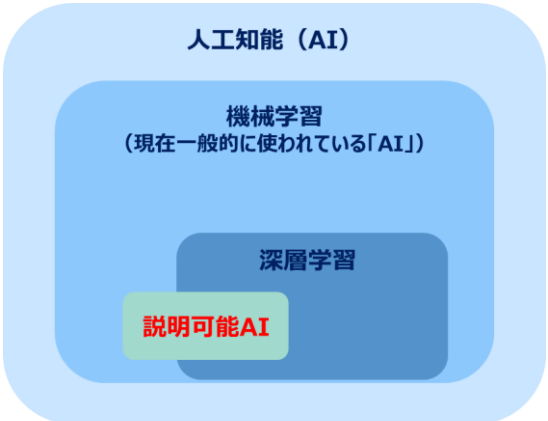

図1　AI、機械学習、深層学習と説明可能AIの関係
出所）各種情報をもとにARC作成

ている。

一方で、深層学習は、推論の過程がブラックボックスになっており、なぜそのように予測をしたのか、判断の根拠を人が理解できないという問題点がある。推論の過程が解釈しやすい機械学習はホワイトボックス型AIと呼ばれるが、一般に予測精度はブラックボックス型AIに劣る。予測精度が高く、かつ根拠を説明できる、説明可能AIが求められている。

なお、各入力データが予測結果にどの程度影響を与えたかを「説明性」、機械学習の推論プロセスが理解しやすいかを「解釈性」という。

◆医療や自動運転などの、AIを活用したサービスや開発で求められる説明性

例えば自動運転における歩行者の検知や走行ルート計画において、AIの誤った予測は事故などの重大な問題につながる。しかしブラックボックス型のAIでは、予測の根拠を人間が理解できずエラー要因の解明や、学習モデルの改良が難しい。病気の診断を支援する場合、AIの判断根拠を説明できないと、医師が診断の確定や患者への説明、投薬の判断などを実施することが困難になる。製造業で予知保全や材料選定を行う場合、異常検知や材料配合比などの予測結果だけでは十分な洞察ができず、製造工程や開発の効率化にはつながらない。

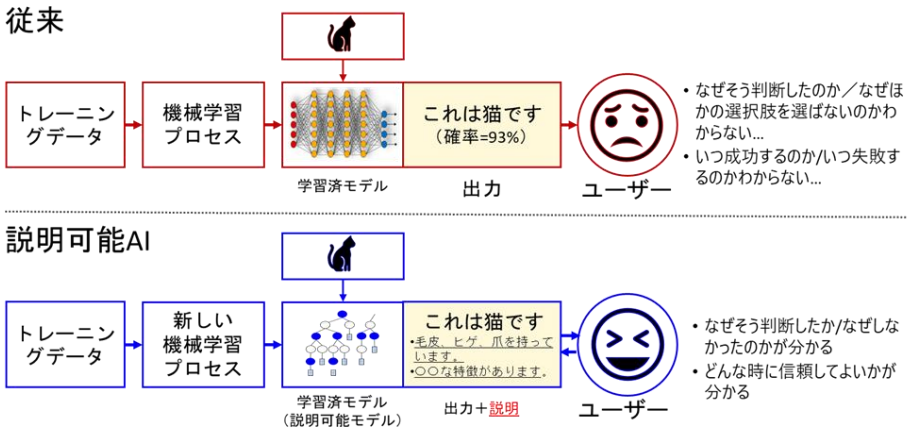

図2　従来のAIと説明可能AIのイメージ

出所）DARPA資料をもとにARC作成

一方で説明可能AIを利用した場合、自動運転の開発ではAIがエラーを起こす環境や要因などを人が理解できるため、事故時の原因究明だけでなく、エラーを事前に回避する学習モデルやトレーニングデータの学習によって、自動運転の精度向上につなげることができる。

医療の診断支援においても、AIの診断根拠を、疾病箇所の色付けや定量的なデータで示すことで、医師の説明責任を助け、患者の納得感を得られやすくなる。

製造業では、故障や材料物性などに影響している入力データを、オペレーターや開発者が理解することで、既知の情報と組み合わせて新たな知見を見出し、製造工程の改善や新たな開発のアプローチにつなげられる可能性がある。

◆規制・倫理面からAIに説明性を求める動き

18年に施行された欧州の「一般データ保護規則（GDPR）」は、22条で「自動化された意思決定」についての透明性（説明責任）を要求しており、意思決定をするAIのブラックボックス化を禁止した。また19年4月には欧州委員会が「信頼できるAIのための倫理指針」を制定した。

表1　各国・世界で議論・策定されるAI社会原則

日本政府「人間中心のAI社会原則」
欧州委員会「信頼できるAIのための倫理指針」（Ethics Guidelines for trustworthy AI）
IEEE「倫理的に配慮されたデザイン」（Ethically Aligned Design）
OECD「人工知能に関するOECD原則」（OECD Principles on Artificial Inteligence）※42か国が署名

出所）　2019年度 人工知能学会全国大会「機械学習における説明可能性・公平性・安全性への工学的取り組み」をもとにARC作成

国内では19年3月に、政府が「人間中心のAI社会原則」を発表し、「人間中心の原則」や「公平性、説明責任及び透明性の原則」が示された。19年7月、総務省が発行した「AI利活用ガイドライン」では、AI利用において、AI　の判断結果の説明可能性や、AIからの説明を前提とした人間の判断の実効性、を求めている。

19年11月に、スタンフォード大学がまとめたAI年次レポートによると、対象とした59のAIと倫理の原則に関する文書の中で、最も頻繁に言及されている倫理的課題として、公平性、解釈可能性、説明可能性の3つがあげられている。

◆説明可能AIのアプローチ方法

機械学習は、一般に予測精度と説明性がトレードオフの関係になることが多い。例えば予測精度が高い深層学習などのブラックボックス型AIは説明性が低く、決定木や線形回帰などのホワイトボックス型AIは予測精度が低くなる傾向

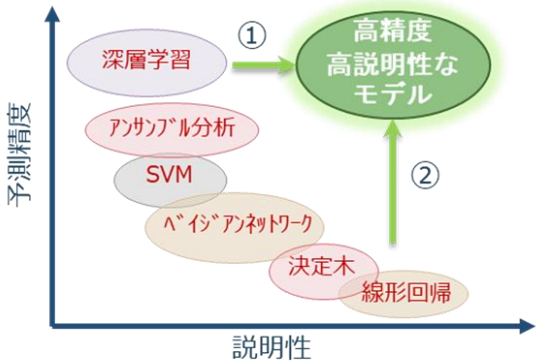

図3　機械学習モデルの精度と説明性

出所）DARPA資料ほか各種公開資料もとにARC作成

にある。説明可能AIは、高い予測精度と説明性を両立する機械学習モデルであり、開発のアプローチは大きく分けて2つある。1つはブラックボックス型AIに、なぜそのように予測したか説明する機能を加えるアプローチ（図中①）である。もう1つはホワイトボックス型AIを発展させ、高い精度を目指すアプローチ（図中②）である。それぞれの機械学習モデルには長所・短所があり、すべての分野に適した機械学習モデルはない。そのため、各機械学習において説明性と精度を両立させる手法が検討されている。

◆深層学習の予測内容を説明する技術

深層学習のようなブラックボックス型の手法に説明性を加える①の技術としては、前述の不正検出のように、特定の入力に対する判断根拠を示す「局所的な説明」の研究が特に活発に行われている。代表的な手法としてLIME、SHAPがある。

１）LIME（Local Interpretable Model-agnostic Explanations）

LIMEは予測結果を線形回帰で近似し、その係数から入力データの影響度を説明する。ただしモデル全体の線形回帰は不可能なので、対象とするデータの周辺からサンプリングを行い、局所的な線形回帰を行う。例えばネコの画像分類では、「ネコ」と判断した局所的な領域のみを表示することで、判断根拠を示す。

２）SHAP（SHapley Additive exPlanations）

SHAPは予測結果に対する入力データの寄与度を定量化し、入力データがプラスに働いたのか、あるいはマイナスに働いたのかなどを説明する。図5の例では、ある予測モデルの結果に対し、年齢、血圧、BMIがプラスに、性別がマイナスに働いたことを示している。

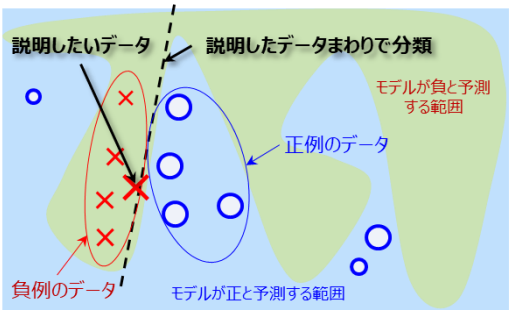

図4　LIMEによる分類のイメージ
出所）arXiv：“Why Should I Trust You?” Explaining the Predictions of Any Classifier をもとにARC作成

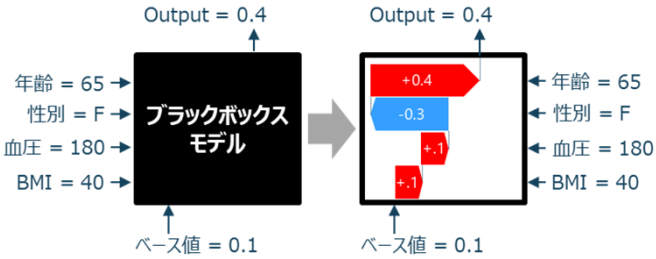

図5　SHAPによる特徴数の定量化イメージ
出所）https://github.com/slundberg/shapをもとにARC作成

◆説明可能AIの開発や応用事例：フェイクニュースや発がんリスクの説明

産業面での必要性やGDPRなどの社会的な要請を受け、説明可能AIの研究開発が活発化している。説明可能AIに関するサーベイ論文によると、関連論文の数は16年以降に急増し17年には100件以上の論文が投稿されている。

17年には米国の国防高等研究計画局（DARPA）が5ヵ年のXAI（Explainable AI）プロジェクトを開始した。実行中のプロジェクトとしては、各種情報からフェイクニュースを特定し、フェイクだと判断した箇所を表示する取り組みや、自動運転中の走行ルート判断をヒートマップと文章で説明する取り組みなどがある。

国内では、19年度のNEDO事業「次世代人工知能・ロボット中核技術開発/人工知能の信頼性に関する技術開発」において、「説明できるAI」関連のテーマが7件採択されている。横浜国立大学らの研究チームは生体データを用いて発がんリスクを説明できる「高信頼性進化的機械学習」の開発に取り組んでいる。個人のマイクロRNAの状態と発がんリスクとの関係性を説明可能AIによって示し、食生活の提案などのヘルスケアサービスにつなげる。

表2　人工知能技術の説明性に関する研究開発（「説明できるAI」）採択テーマ

採択テーマ	委託先
生体データを用いて発がんリスクを説明できる"高信頼性進化的機械学習"の研究開発	東京医科大学、キユーピー、横浜国立大学
視覚的説明と言語的説明の融合によるXAIの実現に関する研究	中部大学、情報通信研究機構
モジュール型モデルによる深層学習のホワイトボックス化	東京工業大学
画像認識AIの誤認識の原因を説明する技術の研究開発	ゼンリン、大阪大学
学習指針をヒトと協創する半自己学習フレームワークおよび知識を創出する情報基盤に関する研究	産業技術総合研究所、BonBon
脳型生成モデルによる推論・言語と正直シグナルの融合によって説明するAIの研究開発とその育児支援への応用	大阪大学、電気通信大学
臨床現場での意思決定を支援する人工知能基盤の開発	サスメド

出所）　NEDOホームページをもとにARC作成

AIを用いた材料開発技術のマテリアルズインフォマティクスにおいて、解釈性の高いモデルを使い、新たな科学的知見を見出しながら材料開発を進める動きがある。19年10月、NECらの研究チームは、解釈性の高い決定木と線形回帰の組み合わせを応用した機械学習モデルで材料開発を行い、「磁性体でPt原子のスピン分極率の総量が大きい材料は熱電効率が高い」という新たな知見を見出した。この知見から熱電効率の優れた材料を推定した。推定した材料を合成し評価した結果、従来よりも熱電効率に優れた材料であることが明らかとなった。

◆説明可能AIの実用化例：　AIサービスへの実装化、金融審査などで活用

19年以降、ITベンダーが自社のAIサービスに、説明可能AIの機能を実装する動きが拡がっている。19年11月、グーグルはクラウド型AIサービスに

「Explainable AI」と呼ばれる説明可能AIの機能を付与した。例えば画像分類の予測結果に対し、入力データの中で強い影響を与えた領域を色付きで表示する。DataRobotも自社のAIサービスに、「特徴量のインパクト」と呼ばれる、予測結果に対する各入力データの影響度を定量化する機能を付与している。

日立は住信SBIネット銀行と「AI審査サービス」を運用している。予測結果の根拠を定量化する「影響度算出技術」を用いて、審査における家族構成や年収などの影響度を算出し、顧客への説明に利用している。20年1月には説明可能AIを活用した、業務システムへのAI導入を支援するサービスの提供を開始した。

NECは、解釈性の高い決定木と回帰分析を組み合わせた「異種混合学習」と呼ばれる手法を活用している。対象データを決定木で複数に分割し、それぞれについて回帰分析をすることで、高い精度と解釈性を両立させる。19年12月には、深層学習と異種混合学習を組みわせた不公正取引の審査業務を支援する「AI売買審査支援サービス」の提供を開始した。

富士通は、入力データから仮説を列挙し、暗黙知を顕在化させる「Wide Learning」と呼ばれる手法を開発している。19年9月に機能を拡張し、有望顧客の特定などの従来機能に加えて、マーケティング施策や機械の制御方針などの、最適な行動計画を提案するサービスを開始した。

◆期待が高まる説明可能AI、AIの利用目的を踏まえた必要性の判断も重要

説明可能AIは、予測以外の「ユーザーが必要とする情報」を「見える化」する技術である。何が「ユーザーが必要とする情報」か、どう「見える」ようにするかは、ユーザー自身が明確にする必要がある。また、AIの利活用においてブラックボックス型であることが障壁となる場合の、やむを得ない解決手段である。例えば、ゲーム用AIや巡回ルートの判断支援AIなど、AIがブラックボックス型であっても大きな問題がない場合に、説明性を付与する必要はない。

説明可能AIは、現在のAI開発やAI利用におけるブラックボックス問題を解決し、さらなるAIの普及拡大を支える重要な技術であり、今後もさまざまな要素技術の開発やサービスへの実装が期待されている。一方で、説明可能AIの必要性は、AIの利用目的をよく理解したうえで判断する必要があり、社内のAI人材の育成やAI活用を得意とする企業との連携も重要であろう。　　【塚原祐介】