AI 活用の鍵「説明可能な AI」とは

株式会社第一生命経済研究所　客員研究員　高野　辰也
（第一生命情報システム株式会社　デジタル推進部所属）

（要旨）

○　生活やビジネスのあらゆる局面で AI が使われるようになり、社会は益々便利に変化していく一方、AI の判断根拠を解釈できない「AI のブラックボックス化」が課題として認識されている。そのような中で登場した技術が「説明可能な AI」であり、AI の課題を解決する鍵になるのではないかと、近年注目度を高めている。

○　例えば、国内では内閣府や総務省などが説明可能な AI の重要性を訴えている他、海外では EU の AI 規制案や米国 FTC の規制のように、罰則規定を設けるような例も出始めている。一方で、現在研究されている説明可能な AI は様々な手法が考案されているが、技術的な限界から、説明できることが限られている。

○　このような「″完璧な″説明可能な AI が存在しない中で、企業としては AI の説明責任が求められている」という課題に対して、説明可能な AI でできることを理解し、ビジネス上求められる説明を得るためにどうすればよいかを試行錯誤しながら最適な組み込み方を探り、説明可能な AI を使いこなしていくことが重要となってくるだろう。

1. AI が抱える課題

近年、様々なシステムに AI が組み込まれるようになり、その恩恵にあずかることも増えてきている。例えば、ユーザー観点でみると、通販サイトや音楽動画配信サービスでのレコメンド（推奨）機能や WEB 検索、ビジネス観点でみると、需要予測や異常検知といったところで AI が使われ、サービスの質や業務効率を高めている。

こうした AI サービスが世間に広がっていった背景には、当然、AI の精度向上が寄与している。しかし精度を向上させるために登場した複雑なアルゴリズムやディープラーニングといった手法により、AI の内部処理は複雑になり、その処理や判断過程を人間が解釈することが困難になってきている。このような AI のブラックボックス化は、AI 活用における大きな課題である。

そして、実際にブラックボックス化した AI が組み込まれたシステムを使ったり、提供したりすると、以下のような疑問が生じるだろう。

・　（ユーザー観点）どうしてこの商品・動画が推奨されたのだろうか？

- （ビジネス観点）テストデータでの予測精度は高いが、ビジネス上の妥当な判断を行えているのか？
- （データサイエンティスト観点）構築したAIに不具合がないか？

こうした疑問を解消しないままAIの利用や提供を続けると、以下のような問題に発展する可能性がある。

- （ユーザー観点）予測結果に納得がいかず、ユーザー離れやクレームに繋がる
- （ビジネス観点）ビジネスにマッチした判断がされず、機会損失や売上低下につながる
- （データサイエンティスト観点）テスト時は問題のないモデルが本番運用時に大きく精度が低下する

こうした問題に対し、事前にAIの判断ロジックを読み解き、疑問を解消する技術が「説明可能なAI（Explainable AI、XAI）」である。

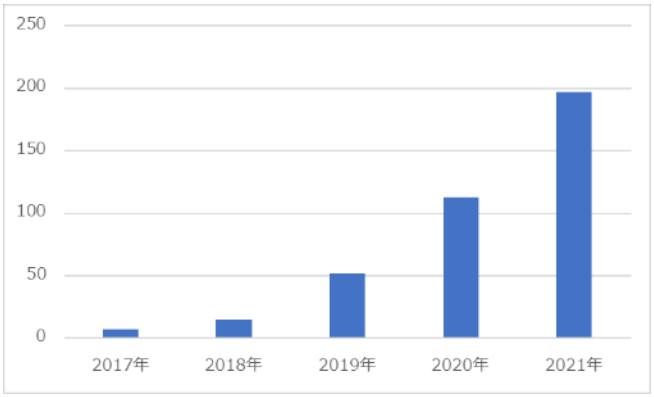
2. 説明可能なAIの動向

説明可能なAIは、近年注目を集めている研究領域であり、検索トレンドや論文数からも、その注目度が窺える。図表1-1は、「Explainable AI」というキーワードの2015年以降のGoogle検索のトレンドを示した図表であり、徐々に注目度が高まっていることがわかる。図表1-2は、arXivという論文検索が可能なウェブサイトにおいて、「Explainable AI」というキーワードで年度別に検索した結果である。ここ数年で、説明可能なAIに関する論文が増加傾向であることが読み取れるだろう。

図表1-1 「Explainable AI」というキーワードの2015年以降の検索トレンド

（出所）Google Trends のデータを元に筆者作成

図表 1-2　論文検索サイト「arXiv」における「Explainable AI」のキーワード検索のヒット数



（出所）論文検索サイト「arXiv」のデータを元に筆者作成

また国内の動向をみると、行政の観点では、内閣府の人工知能技術戦略会議で取りまとめられた「人工知能技術戦略実行計画（案）（2018 年、注 1)」や総務省が公開している「AI利活用原則（2019 年)」の中で、説明可能なAIに関する内容が記載されている。経済界でも、経団連公表の「AI活用戦略（2019 年)」において、説明可能なAIの研究開発の重要性を具体的に述べており、注目が必要な領域であることが理解できる。

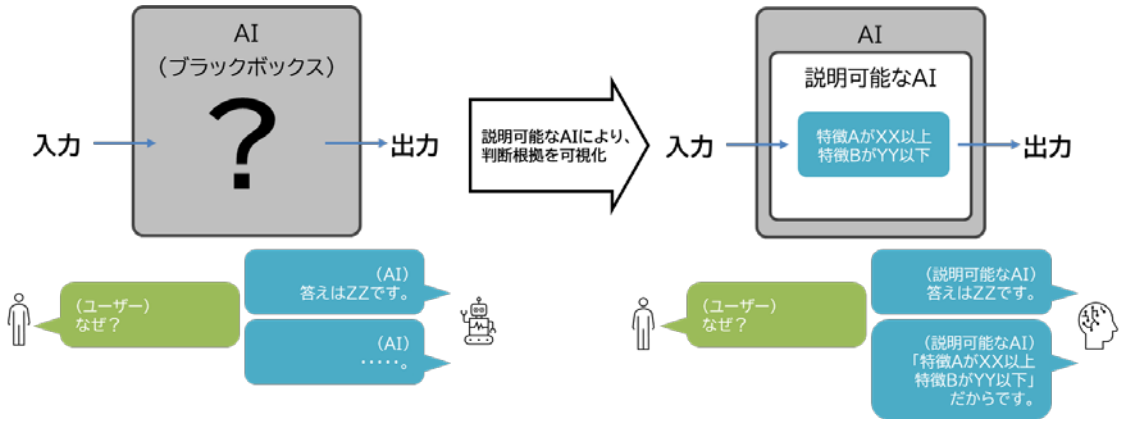
さらに海外に目を向けると、2021 年 4 月にEUが公表したAI規制案では、重要インフラ・教育・採用関連にAIを利用する場合、事前審査によるAIの説明責任を求め、違反企業に対しては罰則を科すものになっている。また米国のFTC（連邦取引委員会）でも、差別につながるAIを開発・使用した企業をFTC法違反として摘発するという声明を発表しており、AIの説明機能の搭載を怠った企業に対して罰金を科すものになっている。このように、違反した場合の罰則まで定められている点は注目すべきと考える。

日本においては、現状は努力目標としての側面が強いが、海外の流れを踏まえ、説明可能なAIなどの技術を用いたAIのガバナンスが義務化される可能性に備えておく必要がある。

3.　説明可能な AI とは何か？

説明可能なAIは、AIの判断に対し、なぜそのような判断に至ったのかを理解できるようにする技術の総称である。図表 2 のように、従来ブラックボックスとされていたAIの判断過程を、説明可能なAIにより可視化できる。

図表 2　説明可能なAIのイメージ



（出所）筆者作成

説明可能なAIは技術の総称であり、実際には様々な手法やアプローチが提唱されているが、「AIモデル全体の挙動を説明する」（＝AIモデルを説明）か、「AIモデルに投入したリクエスト１つ１つの挙動を説明する」（＝予測を説明）のかという切り口で分類されることが多い。AIモデルの開発者としては、まずはAIモデル全体の挙動を把握する必要があるが、利用するユーザーは全体の挙動よりも、個々の予測ごとのより具体的な説明を求めるため、このような手法の差が存在する。

また「特定のアルゴリズムのAIに対して適用できる」ものか、「どのようなアルゴリズムのAIでも適用できる」ものかという切り口で整理されることもある。後者は、インプットとアウトプットをもとに内部処理を説明しようとするアプローチであり、様々なモデルを試しながら精度を高めていくという機械学習プロセスにも適していることから、この手法に対する期待は高い。

このように、一口に説明可能なAIといっても、実体は様々な切り口で多くの手法が日々研究者により考案されているため、説明可能なAIを利用する際は、手法ごとの特徴を捉え、最適な手法を選定していく必要がある。

なお様々な手法のある説明可能なAIであるが、全てのケースで導入しなければならないわけではない。説明可能なAIを実装するにはコストが発生するため、ユースケースによってそれが必要かを判断すべきである。例えば、レコメンドシステムのような間違っていても影響の少ないAIについては不要と判断することもできる一方で、医療や金融などのように間違いが致命的な問題に発展するAIについては、説明可能なAIが必要というように、都度判断していくことになる点は注意したい。

4.　代表的な手法における説明可能な AI の説明内容

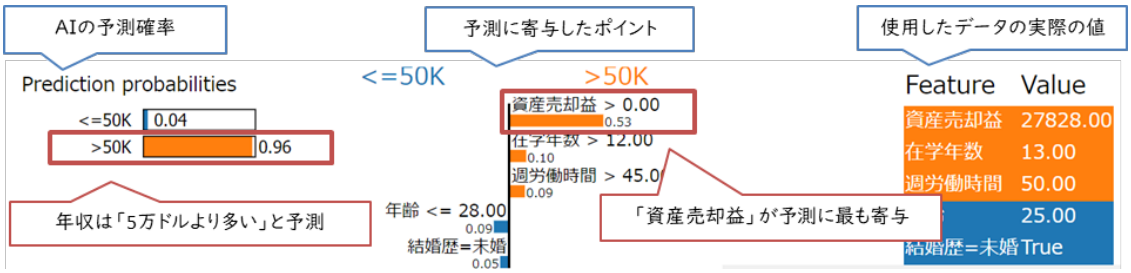
続いて、説明可能なAIにおける代表的な手法であるLIME（Local Interpretable Model-Agnostic Explanations）という手法をもとに、具体的にどのような説明が得られるのかをみてい

きたい。

LIMEは、2016 年の論文で考案された手法であり、説明可能なAIの先駆け的な手法である。LIMEは、AIの予測結果に対して、予測に寄与した特徴量を説明する手法であるが、画像データやテキストデータにも利用できる点や、動作原理としてもシンプルなため、よく使われる手法である。なお、特徴量とはAIのモデルを生成する際に用いられるデータの各項目のことである(図表 3 右端の「資産売却益」「在学年数」など)。

図表 3 は、「年収が 5 万ドルより多いかどうかを予測するAI」の予測結果の説明イメージである。例では、ある一人の顧客について予測を行い、年収は 5 万ドルより多い(>50K)と予測している。資産売却益の大きさが、特に予測に寄与したポイントであることが、説明内容から読み解ける。その他、在学年数や週労働時間なども影響していることがわかるだろう。

図表 3「年収が 5 万ドルより多いかどうかを予測するAI」の予測結果の説明イメージ



(出所)LIMEライブラリ(marcotcr/lime)の実装例を元に筆者作成

また図表 4 については、「ハスキー犬か狼かを分類するAI」において、ハスキー犬をオオカミと誤って予測した画像(左)と予測根拠を抽出した画像(右)である。このケースでは、左側のハスキー犬の画像のうち、右側の画像で灰色ではなく、元の画像が表示されている部分が予測に使用した部分であると説明していることを意味する。つまり背景の雪に着目して狼と判定したことが、説明内容から読み解け、AIが正しく学習できていないことを、説明可能なAIが検知したことを示している。これにより、このAIモデルを構築した際に使用した画像の内、狼の画像の多くは背景が雪の画像だったと推測できる。したがって、背景に雪のない狼の画像を追加して再度AIモデルを作りなおすことで、より正しい判断ができるAIモデルを構築できるだろう。

図表 4 ハスキー犬を狼と誤って予測した予測根拠の説明イメージ

（出所）"Why Should I Trust You?": Explaining the Predictions of Any Classifier(Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, 2016, P9） ※注釈は筆者加筆

上記の通り、現状、説明可能なAIの技術でできることは、「なぜその判断を行ったのか」についての根拠を提示することである。その根拠が妥当なのかの判断は未だ人間が行う必要があるため、実務で利用する際は、説明可能なAIで得られた説明結果を踏まえ、業務上求められる内容や説明レベルに加工して使用する必要がある。

5. ユースケース

では実際に業務で説明可能なAIを適用するとどのようになるのか、医療業界と金融業界の2つのユースケースを例に確認してみたい。

(1) 医療業界におけるユースケース

例えば、レントゲン写真から癌があるかどうかを判定することで、医師の診断をサポートするAIモデルを構築したとする。その際、癌であるという判定が出てもなぜその判定がされたかがわからないと、もちろん使い物にならない。

そこで、前章の図表 4 で触れたような、予測した根拠となるポイントをレントゲン画像にハイライトする説明可能なAIの手法と組み合わせ、医師が集中的に見る（読影する）べきポイントを絞り込んでくれることで、原因特定の効率を高めることができ、有用なAIになるだろう。

(2) 金融業界におけるユースケース

次に貸付審査を判断するAIを例に考えてみる。貸付審査は審査を受ける人の生活に関わることであり、公平な判定が下される必要がある。そしてある顧客に対して、貸付不可の判定を出したとすると、顧客は貸付不可となった理由を確実に問い合わせるだろう。

ここでも、LIMEのような、予測に貢献した特徴を説明する説明可能なAIの手法が必要とされる。前章の図表 3 で触れたような説明内容を得ることで、「貸付残高がXX円以上、前年度年

収がYY円以下のため、貸付不可」という具体的な判定理由が説明できるようになり、金融機関としての説明責任を果たすことが可能となる。

6. おわりに

ここまでで、説明可能なAIの概要や具体的な説明内容、ユースケースなどを説明してきたが、想像していた説明可能なAIと少し違う、と思われた方も多いのではないだろうか。恐らく、そういった感想を持たれた方は、"もっとAIが説明してくれること"を期待されていたのではないかと推測する。しかし前述の通り、現在の代表的な説明可能なAIの技術でできることは、「なぜその判断を行ったのか」についての根拠を提示することである。

一方、第一章で述べたようなビジネス上の課題や、第二章で述べたようなビジネス的な動向や規制からもわかるように、AIの説明責任はもはや必須要件となりつつある。

このように"完璧な"説明可能なAIが存在しない中で、ビジネス側にAIの説明責任が求められていることが、現在の課題であると考える。

この課題を克服するためには、説明可能なAIにできること・できないことを理解し、ビジネス側に求められる説明責任に対して、どのように説明可能なAIを組み込むことで、要件の一部または全部の説明ができるかを試行錯誤し続ける必要があるだろう。また、総務省が公開している「『AIの説明』の現状とこれから(2018 年 11 月)」という資料の中で触れられているように、ビジネスで求められる説明内容と学術的に研究されている説明内容の間に存在する乖離を、双方による一層のコミュニケーションとフィードバックループで解消しようとするアプローチも有効と考えられる。

AI提供に関わる人々がこうした課題解決の取組を繰り返しながら、説明可能なAIがビジネスにおいて真に機能するように社会に組み込んでいくことで、誰もが安心してAIを利用・提供できる豊かな社会が実現することを期待する。

以　上

【注釈】

1) 人工知能技術戦略実行計画（案）は、AI 戦略実行会議の AI 戦略 2019 の根拠となっており、最新の戦略として AI 戦略 2021 が公開されている。

【参考文献】

- Christoph Molnar “Interpretable Machine Learning” (2018) https://christophm.github.io/interpretable-ml-book/
- 大坪直樹、中江俊博「XAI(説明可能な AI)」(2021)
- 内閣府 人工知能技術戦略会議「人工知能技術戦略実行計画（案)」(2018) https://www8.cao.go.jp/cstp/tyousakai/jinkochino/7kai/siryo3.pdf
- 総務省 AI ネットワーク社会推進会議「報告書 2019 別紙 1AI 利活用ガイドライン（4.AI 利活用原則案)」(2019)

https://www.soumu.go.jp/main_content/000637097.pdf
・ 一般社団法人 日本経済団体連合会「AI 活用戦略～AI-Ready な社会の実現に向けて～」(2019)
https://www.keidanren.or.jp/policy/2019/013.html
・ Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin “"Why Should I Trust You?": Explaining the Predictions of Any Classifier” (2016)
https://arxiv.org/abs/1602.04938
・ 総務省「『AI の説明』の 現状とこれから」(2018)
https://www.soumu.go.jp/main_content/000587311.pdf