技術の豆知識

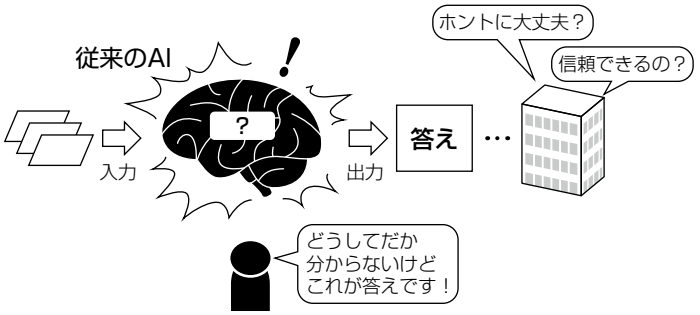
AIを安心して使うために

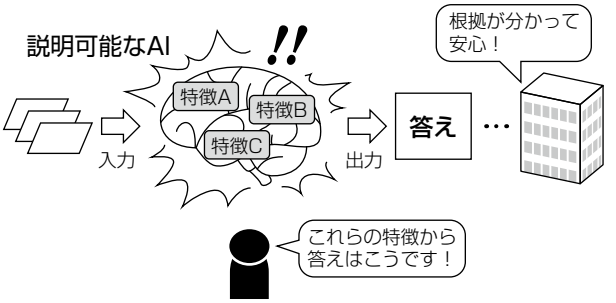
説明可能なAI

ブラックボックスの中身を知りたい

現在，様々なビジネス領域で機械学習が活用されています。しかし，AI(Artificial Intelligence)アルゴリズムが高度になるほど機械学習モデルは解釈が困難な“ブラックボックス”になり，導き出されるアウトプットがなぜそうなるのか，開発者も説明ができません。すると“判断の根拠が説明できないものは安心して使えない”という懸念が広がり，AIが社会での実利用でリスクを持つものと捉えられてしまいます。例えば，AIによる自動運転でもし事故が発生した場合，AIの判断理由が説明可能でなければ原因を探ることができず，導入する企業やユーザーの信頼を得られません。

そこで近年，機械学習モデルをホワイトボックス化する“説明可能なAI(eXplainable AI：XAI)”の研究に注目が集まっています。





XAIの代表的な研究

XAIとは，予測結果や推定結果に至る過程が人間によって説明可能になっている機械学習モデル，又はそれに関する技術や研究分野のことを指します。

XAIの実現を目指して，様々な手法や技術が考案されており，特に2016年以降，機械学習関連の国際会議でAIによる判断の根拠の説明に関する論文が増加しています。

例えば，AIが説明を提示する方法の代表的なものとして，次の四つが挙げられます。

⑴ 重要な特徴の提示

“データのどの特徴が予測・認識に重要だったか”を説明として提示する方法です。例えば，集団を収入の高い群と低い群に分類した場合，収入の高い群に分類された人の重要な特徴(年齢，職種等)を説明として提示します。

⑵ 重要な学習データの提示

“どの学習データが予測・認識に重要だったか”を説明として提示する方法です。例えば，画像認識への影響が強い学習画像は何だったのか，また，ある学習画像が“なかった”としたらどの程度認識精度が変わるかを説明として提示します。

⑶ AIの可読化

予測・認識の過程を“可読な表現で記述する”ことでAIの説明とする方法です。複雑なアルゴリズムを人間が読める簡単な記述に書き換えます。

⑷ 自然言語による説明

“データのどの特徴が予測・認識に重要だったか”を自然言語で説明文として提示する方法です。例えば，画像認識によって鳥の種類を判別したとき，結果とともに“これは，長い首と大きなくちばしを持つ茶色の鳥であるため，カイツブリです。”という説明文を提示します。

XAIの求められる領域

XAIはAIへの不安を解消できる手段の一つとして非常に有効なものです。

冒頭に挙げた自動運転のほかにも，裁判，医療診断や予防保全など，大きな信頼を求められる分野では特に判断の根拠について説明が必要です。AIが判断した過程について，人間が理解可能な説明があれば，その判断／処理結果を採用するか否かの指標にすることもできます。

これからの社会でAIの活用領域を広げるために，XAIの早期実用化が求められています。

参考文献：総務省：「AIの説明」の現状とこれから　https://www.soumu.go.jp/main_content/000587311.pdf