第３章

説明可能性　インプットとアウトプットの間にある領域

あなたは理想通りの家を見つけ、その足で住宅ローンを組もうと銀行に向かう。融資担当者の前に座り、氏名、生年月日、職歴など、求められた情報を申込書に記入する。さらにクレジットカードの明細書と給与明細、行っている投資のポートフォリオに関する情報などを提出する。融資担当者は受け取った紙の束を、コピー機のようなマシンの片側から投入する。マシンは瞬時に紙を吸い込み、ウィーンと音を立てると、１枚の紙をプリントアウトする。そこにはこう書かれている――却下。

「申し訳ございませんが、住宅ローンのお申し込みはお断りさせていただきます」と、融資担当者が告げる。

理想の家を見つけたというあなたの興奮は、悲しみと混乱に変わる。「どうして？」とあなたは尋ねる。

「マシンがそう判断したからです」と担当者は答える。そしてもっと説明しろと言うあなたにこう告げる。「お客様にご提出いただいたデータをすべてインプットしました。マシンはそれを他のお客様のデータと比較したのです。承認された人、却下された人、貸し倒れになった人、返済した人と比較したところ、お客様のデータは明らかに貸し倒れになった人のものと似ています」

この時点で、あなたの混乱は怒りに変わっている。「自分の申請のどこが悪くて却下されたのか知りたい。ふざけないでくれ……説明しろ！」

しかし激しい怒りもむなしく、あなたは行き詰まる。融資担当者が説明を拒否しているわけではない。説明できる人物を出すのを拒否しているわけでもない。誰も説明できないのだ。マシンはブラックボックスで、どのようにアウトプットが行われるのか、中をのぞくことはできないのである。

このような状況を受け入れがたいと感じるのは、あなただけではない。しかしこれは、多くの機械学習（ML）アルゴリズムで発生している状況だ。多くの場合、アルゴリズムを開発した人でも、そのアウトプットを説明することはできない。そしてより多くの人々が、MLのアウトプットを説明できるようにせよと要求するようになっている。それがブラックボックスであることを拒否し、透明な箱になることを望んでいるのだ。

問題は、顧客や従業員らに不満をもたらすことだけではない。場合によっては、住宅ローン審査の結果など、決定に対して説明をすることが法的に義務付けられていることもある。今後、ますます多くのAIによる意思決定に対して、同様の義務が生じると覚悟しておいた方がいい。決定の対象となった人々の前ではなく、あなたの開発したAIに差別されたという訴えを起こした人々の弁護士の前に立たされたとき、住宅ローンや面接、保護観察を却下した理由を説明したい、説明できないと、とあなたは思うはずだ。他にも従業員から、「なぜ私は昇進できなかったのか」、「なぜ欠員のあるポジションに自分が就けなかったのか」といったことに説明を求められるケースもあるだろう。そのような場合に説明できないでいることは、倫理的リスクであると同時に、評判・規制・法的リスクにもなり得る。

AIにおける説明可能性の議論は、一般的に、非常に狭い範囲に限られている。バイアスの場合と同様、議論の中心となっているのは、ブラックボックスの中をのぞくためのさまざまな技術的アプローチだ。このアプローチは非常に難しく、場合によっては不可能とさえ言われている。しかしそうした

技術的アプローチに注目するのではなく、またすべてのAI倫理に関する声明に「説明可能性」や「透明性」という言葉をつけて、会議の講演者たちにブラックボックスの存在を非難させる前に、一歩下がって大局を見よう。なぜなら、これから見ていくように、MLのアウトプットが説明可能であることは必ずしも重要ではなく、また説明において、インプットとアウトプットの間で起こっていることは必ずしも必要ではないからである。

説明を分解する

先ほど取り上げた、理想の家を買う夢が破れたという例を再び考えてみよう。そうなるまでに、いろいろなことが起こっている。

1. 少し前、自動化の時代が始まる以前は、銀行の該当部署は住宅ローン申請書のフォーマットを用意し、申請者に対して自分に関する情報を書き入れるよう求め、その情報に基づいてローンを承認すべきか否かを判断する、簡単なディシジョンツリー（決定木）を作成していた。
2. この申請書は、長年にわたり、さまざまな人々によってさまざまな形で更新されてきた。
3. 銀行は、どのローン申請者が住宅ローンを滞納したか、あるいは無事に完済したかを追跡している。
4. 少しすると、何万件という申請が承認され、却下される。そしてどのローンが完済されたか、あるいは不履行になったかを、銀行は把握する。
5. MLを使って融資審査を自動化するのは良いアイデアだと、誰かがゴーサインを出す。
6. 銀行は手元にある過去の情報をすべて利用することにする。また彼らは、申請に関係すると考えられる他の情報も集められることに気づい

ている。たとえばソーシャルメディアのデータ（申請者がどんなソーシャルメディア・サイトを使っているか、そこでどのくらいの頻度で投稿や他人の投稿へのコメントを行っているか、誰の投稿にコメントしているか）などだ。

7. 銀行のチームは、このタスクに適していると思われる学習アルゴリズム（すぐに利用可能なものが多数ある）を選択する。
8. 選択されたアルゴリズムは、膨大な量のデータを解析し、数千のデータポイントからパターンを見つけ出すことに長けている。
9. MLは、申請者ごとに住宅ローンが不履行になる確率を0から1の間でアウトプットする。たとえば0.3345は、不履行が起きる確率が3分の1より少し上であり、0.0178なら2パーセント弱、といった具合である。
10. チームは、不履行の確率が3.74パーセントより高い人に対して、住宅ローンの申請を却下すべきであると決定する。つまり3.74パーセントは承認すべきか否かの基準である。
11. 銀行の経営陣はAIの導入を承認する。
12. 融資担当者のマーヴィンが、あなたの申請が却下されたことを告げる。
13. あなたの弁護士が、汗びっしょりで宣誓証言を行う主任技術者をにらみつけ、なぜ住宅ローンを拒否されたのかの説明を求める。
14. また、あなたは黒人だ。

これだけ端折った説明からでも分かるように（皆さんが考えもしないような判断ポイントをこの中にもっと追加できた）、なぜあなたのローン申請が却下されたかということについて**どんぴしゃり**の説明はあり得ない。というより、無数の出来事からなる1つの大きな説明があり、そしてその無数の出来事に対して、小さな説明が存在しているのだ。要するに、こんな具合だ。ローン申請書を最初に作成した融資担当者は、審査基準をどのような理由で

選んだのだろうか？　なぜその基準は何年もかけて更新されてきたのか？　その更新の妥当な理由として、どんな問題やきっかけがあったのか？　エンジニアやデータサイエンティストは、なぜソーシャルメディアのデータが関連するかもしれないと考えたのか？　なぜ他のデータ、たとえば通っていた小学校のデータが関連すると思わなかったのだろうか？　どうやって学習アルゴリズムを選んだのか？　なぜモデルは、与えられたインプットに対してそのようなアウトプットを行ったのか？　なぜ3.74が閾値なのか？　なぜ3.76パーセントや12.8パーセントではないのか？　経営陣はどのような根拠でAIの導入を承認したのか？　こうした質問（他にも考えられるかもしれない）に対する回答が集まって、あなたがローンの審査で落とされた理由の大きな説明となるのである。

これらの質問と、それに対する答えがすべて揃ったところで、さらに2つの質問がある。

1. 人々がMLのアウトプットの説明を求めるとき、彼らは何を求めているのか？
2. そうした説明の中で、何が最も重要なのだろうか？

ブラックボックスを解明する

住宅ローン申請を却下された理由についての説明の大部分は、「人々がなぜそのような決定をしたのか」を理解するための内容となっている。これを「人間による説明」と呼ぶことにしよう。

人間による説明がどのようなものかは、見当がつくはずだ。「自動化することにしたのは、申請の数が処理できる限界を超えつつあったからです。ソーシャルメディアのデータに関連性があると考えたのは、そこからローンの返済を予測する行動パターンを明らかにできるかもしれないと判断したから

です。3.74パーセントを閾値として設定したのは、アウトプットのクラスタリングの方法と、当行のリスクアペタイト〔ある組織において、受け入れられるリスクの種類や量を示したもので、それぞれの組織内でリスク分析を行った上で設定される〕を組み合わせて判断した結果です」といった具合だ。さらなる説明を求めることもできる。「単に応募数に上限を設けるか、申請を処理する担当者を増やせばいいのでは？　ソーシャルメディアのデータに関連性があるかもしれない、などという賭けに出たのはなぜか？」などのように。

その一方で、「マシンによる説明」というものがある。これは少し奇妙な存在だ。上記の8番目の説明に関連するもので、私たちが求めているのは、モデルがどうやってインプットからアウトプットを導き出したのかという説明である。これについて、念頭に置いておかなければならない質問が2つある。

- インプットをアウトプットに変換するルールはどのようなものか？
  - インプットに使用する大量のデータがあるとする。MLモデルはそのデータを取り込み、データの中に存在しているさまざまなパターンに気づいて、アウトプットを行う。たとえば、あなたの愛犬ペペがどのような外見をしているか、1000枚の画像を使ってMLを学習させたとしよう。MLはそれぞれの画像をピクセル単位で分析し、たとえば373番のピクセルがどうなっているか、そのピクセルと他のピクセル（そして他の何千ものピクセル）の間にどのような位置関係があるかを分析し、あなたの犬がどのように見えるか学習した。たとえば犬が座っている画像では、このピクセルが他のピクセルに対してこういう関係になる、あるいは立っている画像では、別のピクセルが他のピクセルに対してこういう関係になる、といった具合である。つまり「ピクセルがこういうふうになっているとき、それはペペの画像であり、そうでなければ、ペペの画像ではない」という大まかな「ルール」を学習するのである。マシンによるこの種の説明を、「グローバル説明」と呼ぶ。

- これらのインプットを与えた場合に、なぜこのようなアウトプットが得られたのか？
  - なぜ固有のプロファイルを持つあなたは、住宅ローンの申し込みを却下されたのか？　転職を繰り返していたからだろうか？　5年前に無謀運転という軽犯罪で起訴されていたのが悪かったのか？　クレジットカードを使い過ぎたから？　ある男性のSNSに大量のコメントを書き込んだことが原因か？　この種の質問に対するマシンによる説明は、「ローカル説明」と呼ばれる。

MLは複雑なパターンを認識してくれる。実際、人間の理解を超えるほど複雑なパターンでも処理できる。たとえばあなたは、画像に含まれるピクセルの数と、それらが他のピクセルとどのような数学的関係にあるのかを理解して、ある画像を「これはペペ」もしくは「ペペじゃない」と判断できる法則を見つけることができるだろうか？　あるいは特定の画像をMLが「これはペペ」と判断した理由を理解することは？　そんなのは無理、の一言だ。

これが人間による説明と、マシンによる説明の違いである。私たちは人間による意思決定について、私たちが理解できる言葉で説明する。私たちは、その説明で明確にされた関係性を把握できる。ローンの申し込みが殺到したら、その問題に対して、自動化を含むさまざまな解決策が検討されるようになることを理解できるわけだ。しかしマシンによる説明はというと、それは私たちにとって極めて複雑なものである。そこに登場する変数の数と、それらの間に存在する関係性の数の両方が、人間の貧弱な頭脳を混乱させる。たとえ、この複雑さを表す数学的言語を概ね理解できたとしても、である。

人が説明可能なAIを求めるとき、彼らは何を求めているのか、これで理解できただろう。それは人間による説明か、マシンによる説明か、あるいはその両方である。そして人がマシンによる説明を求めるとき、彼らはグローバル説明、ローカル説明、あるいはその両方を求めている。

いっそのこと、すべての説明を常に行うのがよいのではないかと思うかもしれない。しかしマシンによる説明を手にするには、それなりのコストがかかり、他にリソースを割きたいこともあるだろう。最も重要なのは、説明可能なモデルという目標を達成するためには、精度の低下という代償を払わなければならないことが多いという点である。なぜそうなるのかというと、簡単に言ってしまえば、まさにMLの精度を上げるための特徴が、私たちの理解度を下げるからだ。その特徴とは、MLが理解するパターンの複雑さである。他の条件を同じにした場合、データが多ければ多いほど、MLはより多くの（複雑な）パターンを認識できるようになり、その結果、精度が向上する。別の言い方をすれば、他の条件が同じであれば、学習する例が多ければ多いほど良いということである。しかしMLがより多くのデータとより多くの（複雑な）パターンを見つければ見つけるほど、何が起きているのかを理解できる可能性は低くなる。説明可能性を向上させれば、精度は低下する。その逆もしかりだ[(1)]。

このことは、私たちに新たな「コンテンツからストラクチャーを導く教訓」を示している。

コンテンツからストラクチャーを導く教訓5

特定のユースケースにおいて、人間による説明、マシンによるグローバル説明、マシンによるローカル説明のどれが重要なのか、もしくはそのすべてが重要となるかの判断をする、適切な人物が必要。

そうした人物が（どの説明がいつ重要となるかをどう判断するか）検討する上で参照すべきことは、そもそもなぜ説明が重要なのかということに左右される。

説明の重要性

あなたは結婚している。自分で思う限り、妻との関係はそれなりに上手くいっているようだ。DVはなく、笑いがあり、親密な時間も少なくない。しかしある日、あなたが目を覚ますと、配偶者が荷物をまとめて出ていこうとしている光景が目に入ってきた。

「何してるの？」とあなたは尋ねる。
「別れましょう。ペペは連れていくから」
「どうして？」とあなた。
「別に」と彼女は答え、犬を従えてドアから出ていく。

あなたは怒り出す。何しろ、住宅ローン申請を却下した銀行を訴えて勝訴し、夢のマイホームに引っ越したばかりなのだから。しかしその広い室内は、あなたの孤独を嘲笑うかのようだ。

深夜3時、カリフォルニアキングサイズのベッドの上で横になるあなたの頭に、「どうして？」、「どうして彼女は自分を捨てたんだ？」という疑問が何度も浮かぶ。この疑問への答えを求めるのには、少なくとも3つの理由がある。

第1に、何の説明もないというのは失礼だ。あなたは投げ捨ててしまっていいような物ではない。あなたは価値のある人間であり、その価値を示すものとして、説明を受ける権利がある。説明を拒否して立ち去るのは、傷口に塩を塗るような行為だ。

第2に、もし説明があれば、それに対して何か手を打てるだろう。彼女が出ていく理由は、あなたが自分に関心を持っていないと感じたからではないか？　それならば、家に帰ったら携帯電話の電源をオフにすると約束したら、彼女の気が変わるかもしれない。それともロマンティックな雰囲気が足りなかったのか？　それならば、もっとロマンティックになれるように頑張ることができる。ロマンスについてアドバイスしてくれる人はいるだろうか？　ググってみよう。それとも住宅ローンを払うのに手いっぱいで、苦しい暮らしをさせていたからだろうか？　それならばいっそ家を売ってしまって……ちくしょう、どうせあの銀行は最悪だったし、不履行にしてマーヴィンが正しかったことを証明してやろうか。それで彼女と一緒にコスタリカへ逃げることだってできる。

第3に、共同生活していく上での一般的なルールを知り、そのルールが対応可能なものかを考えたい。自分に何が期待されているのか？　仮に今日、何かひとつ解決できても、明日には別の問題が起きるのだろうか？「ここではどのようなルールに従わないといけないんだ？　私の母についてできることは何もないから、もしそれが関係しているとしたら、本当に不公平だ。あるいは母のせいではないのかもしれない……けっきょく『育ってきた環境が違う』って言ってたのはこのことなのだろうか。最初からずっと『私は白人だけどあなたは黒人』という関係性があったのか？」

これら3種類の懸念は、MLのアウトプットにおいても起きる可能性がある。融資や面接を受けられる人を判断したり、表示する広告を決めたり、マッチングアプリで誰にどのユーザーを紹介するかを決めたりするMLがあったとしたら、あなたはその決定の裏側にある説明を求めるだろう。それは敬意を示すものだからであり、決定を変えたいと思ったら何ができるかを考えるのに役立つからであり、また押し付けられたルールがそもそも公平なものかどうかを判断したいと思うからである。融資を断られたのは、自分が黒人だったからなのだろうか？

しかし、説明可能性の重要性を示すこれら3つの理由は、私たちに明白な指針を与えてくれるわけではない。忘れてはならない――説明可能なMLを実現するにはコストがかかる。説明可能性の重要性と、精度のような他に考慮すべき事項、そして説明可能性の実現に割くことのできる（あるいは足りない）リソースとの間でバランスを取る必要があるのだ。場合によっては、マシンによる説明は必要ないという判断をするかもしれない。またマシンによる説明は「必要」ではなく「あればよい」程度だと考えるかもしれない。一方で、それが必要不可欠な場合もある。

このような判断と同様に、それぞれのユースケースにおいて説明可能性がどのていど重要かを判断するための単純なディシジョンツリーは存在しない。しかしその重要性が分かっていれば、どのように検討すればいいのかを理解できる。次に、マシンの説明可能性が問題になる場合と問題にならない場合を、それぞれいくつか示して解説しよう。これらは厳密なルールではなく、経験則であることがお分かりいただけるだろう。

マシンの説明可能性が問題にならない場合

人がどう扱われるべきかについての決定を、開発したモデルが直接的に行わない場合

玩具の工場に出荷するネジの納期を予測するために、MLを活用するというケースを考えてみよう。この場合はおそらく、それほど大きな倫理的リスクはないだろう。予測するのは人間に関する事柄ではなく納期であり、またその予測が間接的に誰かの扱いを悪くすることにつながったとしても（遅延があった場合に誰かが非難されるなど）、サプライチェーンに関する予測は、倫理的リスクを本質的に抱えるものではない。ここで気にすべきは精度であり、マシンの説明可能性を優先させる必要はないと、合理的に判断できる。ここで説明可能性が重要だと考えられるケースは、たとえば「説明可能なモ