保険・年金フォーカス

不公正バイアスへの対処

保険事業への不公正バイアスの混入をどう避けるか?

保険研究部 主席研究員　篠原　拓也
(03)3512-1823　tshino@nli-research.co.jp

1――はじめに

現在、AI(人工知能)の開発・活用が、社会のさまざまな分野で進んでいる。保険分野でも、見込み客の選定やアプローチ方法の検討(マーケティング)、商品設計、加入時の引受査定や給付時の支払査定、保有契約の管理など、多くの面で、AI の導入が進められている。

通常、AI は顧客や経営環境等に関する膨大なデータをもとに、経営の判断や、事業の予測に役立つ結果を計算して、出力する。その際、現実の実務等を模したモデルを用いることが一般的だ。

こうしたモデルを通じた AI の活用は、人間が行ってきた従来の作業の質を、時間、費用、正確性などの面で、飛躍的に向上させる。しかし、AI が出す答えが常に正しいとは限らない。AI が倫理的な面から問題のある答えを出す―そんな事例が、最近、しばしば浮かび上がっている。これは、「不公正バイアス(unfair bias)」と呼ばれるもので、その動向が注目を集め始めている。

本稿では、保険事業に混入する恐れのある不公正バイアスと、その対処法について、海外の文献をもとに見ていくこととしたい。

2――不公正バイアスとは

はじめに、不公正バイアスとはどういうものか、おさえておこう。

1｜不公正バイアスの定義は包括的

まず、不公正バイアスとはどういうものか、その定義をおさえておこう。アメリカのアクチュアリー会(SOA)の研究機関が公表したペーパー[1]によると、不公正バイアスは、“unexplained adverse outcomes for marginalized communities”(非主流とされるコミュニティにとって、説明のつかない不利益な結果[筆者邦訳])と定義される。何か漠然としている感はあるが、さまざまな事象に対して包括的な定義を行おうとすると、このような形になるものと思われる。

不公正バイアスの基準は、時とともに変化していく。例えば、保険契約を含めて、さまざまな契約の申し込みでは、利用者が自分の性別を書類に記入したり、画面上で入力したりするケースがある。

[1] “Avoiding Unfair Bias in Insurance Applications of AI Models”(SOA Research Institute, Aug. 2022)より。

かつては、その区分は「男性」と「女性」だけだったが、近年は「その他」や「回答したくない」なども見られる。これは、トランスジェンダー等の“非主流とされるコミュニティ”にとって、「男性」か「女性」かを二律背反のように選択することは、“説明のつかない不利益な結果”をもたらす、との認識が社会に広がってきていることのあらわれと見ることができる。

2｜不公正バイアスは AI 以前から存在していた

不公正バイアスは、AI の普及とともに新たに生じた問題ととらえられがちだ。だが、この問題は、AI 導入以前から存在していたもので、元をただせば人間に依拠しているものと言える。一般に、AI は膨大なデータをもとに機械学習を行い、モデルによる判断や予測の精度を高めていく。その際に用いるデータのなかに、不公正バイアスの芽が入り込んでいる可能性がある。そうしたデータを機械的に処理して、モデルを向上させていくうちに、不公正バイアスも拡大していくこととなる。つまり、不公正バイアスは、人間がもともと持っていた差別や偏見に根差していると言えるだろう。

このため、不公正バイアスに対処するためには、AI が構築するモデルのリスク管理だけではなく、それを取り扱う人間の道徳や倫理の教育が必要となる。

3｜不公正バイアスは幅広い業務に潜在している

不公正バイアスの発生は、直接顧客に対応する場面だけにとどまらない。リスクは、マーケティングや商品設計、事務管理、広告・宣伝など、幅広い業務に潜在している。AI システムでは、モデル設計が不適切であることから生じるリスクは、「モデルリスク」と言われる。先ほどの SOA の研究機関のペーパーによると、AI に混入する不公正バイアスについては、モデルリスクの管理と統合して対処することで、効率的な業務運営が実現できるとされている。

3――保険事業での不公正バイアス混入

不公正バイアスは、保険事業にも混入する。具体的にどういうケースがあるか、見ていこう。

1｜顧客との接点は、常に、不公正バイアスのリスクにさらされている

保険事業は、目に見えない保障の内容を、保険契約者と保険会社の間で共有することが前提となる。一般的には、営業職員や保険代理店が顧客に保障内容を説明して理解を促すことが行われる。

AI の観点からすると、近年、ネット専業の保険会社等で行われているチャットボットが、新たな顧客との接点として挙げられる。チャットボットの会話に組み込まれている AI に、不公正バイアスが入り込む可能性は否定できない。

2｜リスク区分の設定には、不公正バイアス混入のリスクがある

保険事業では、保険料や引受条件の設定において、契約者をいくつかの群団に区分することが行われる。例えば、生命保険では、性別や年齢別に保険料を設定することが一般的となっている。こうしたリスク区分は、リスクの多寡に応じた保険引き受けを行うことで、契約者間の公平性を確保する狙いがあると言える。その区分には、客観性や合理性が求められる。それとともに、高い倫理性を満たすことが必要となる。[2]

[2] 例えば、仮に、人種によってある保険の給付支払が異なる、という過去の実績データが得られたとする。この場合、このデ

3｜AIによる支払査定でリスク区分を用いる際も不公正バイアスに留意が必要

近年、保険会社の支払査定プロセスでは、契約引受時のリスク区分をもとに、異なる査定方法を用いる取り組みが行われている。その背景には、AI のモデルを通じた予測手法の確立があるとされる。ただ、こうしたAIによる支払予測にも、不公正バイアス混入の恐れがある。

支払査定プロセスでのAI活用の典型は、給付の不正請求の検出だ。検出を自動化することで、保険会社のリスク軽減に役立つ。一方で、検出に不公正バイアスが入り込む懸念を抱えることとなる。

4――不公正バイアス混入の機序

不公正バイアスはどのように混入するのか。その機序にはいくつかのパターンがある。

1｜保護属性を用いてデータを収集すると、不公正バイアスが混入する恐れがある

データの収集時に、倫理的に不適切な属性を用いることが不公正バイアスにつながる場合がある。こうした倫理的に不適切な属性は、「保護属性（protected attributes）」と呼ばれる。例えば、人種、婚姻状況、宗教、政治的見解などが保護属性として挙げられる[3]。こうした保護属性をもとに、データを収集し、AIを用いて分析を行うと、モデルが出力する結果に、不公正バイアスが入り込むことがある。一般に、保護属性の収集や分析は禁じられており、保険会社はこうした属性に基づいて、モデル結果を分類することはできない。

2｜微細区分によって、不公正バイアスが混入することもある

ある共通の属性に基づいてデータを細かく収集、分析すれば、それをもとに複数の集団区分を設定することができる。こうした区分を用いれば、保険給付支払のリスク属性に基づく、微細集団レベルでの価格設定が可能となる。こうした区分は、「微細区分（micro segmentation）」と呼ばれる。

微細区分に関しては、次のような経路で、不公正バイアスが発生する可能性がある。

(1) モデルが複雑で理解が不十分な場合に、不公正バイアスが混入

大量のデータと多変量リスク評価スコアを用いるモデルは、複雑なものとなりやすい。こうしたモデルでは、得られた結果について、人間の理解が追いつかない可能性が高い。どのような属性がその評価につながっているのか、判断が困難な場合がある。こうした場合に、実務者が意図していないような差別的なバイアスが、結果に入り込んでしまっていることがある。

(2) 保護属性と相関の高い変数を用いることで、不公正バイアスが入り込む

AIモデル内の変数は、保護属性と高い相関関係を有している可能性がある。例えば、アメリカ南部には移民の割合が高い地域があり、地域属性と（保護属性である）人種属性が相関していることがある。地域属性によって微細集団に区分することが、実質的に、人種属性によって区分することにつながってしまう。その結果、保護属性を用いた場合と同様のデータ分析が行われ、不公正バイアスが入り込む可能性がある。

(3) データ取得分野が偏ることで、不公正バイアスにつながることも

ータをもとに、人種ごとに保険料率や引受基準を変える取扱いは、倫理性の観点から問題が生じるものと考えられる。

[3] オーストラリアの労働法制では、人種、皮膚の色、性、性的指向、年齢、身体または精神上の障がい、婚姻状況、家族や介護者による介護責任、妊娠、宗教、政治的見解、国民的出身、社会的見解によって、従業員等に差別的な取扱いをすることが禁じられている。（“Fair Work Act 2009”（Section 351(1), Australia）より） これらの要素が保護属性に該当する可能性がある。

AIの入力データとして、これまで未開拓であった事業分野のデータを用いる場合もある。そうした場合、全般的にデータは不足がちとなる。非主流とされるコミュニティに対するデータが欠落し、データに偏りが生じることがある。その結果、不公正バイアスが生じる可能性がある。

(4) サイバーセキュリティ問題が、不公正バイアスを惹起する懸念もある

一般に、AIを用いたシステムは、サイバーセキュリティの問題にさらされている。不正なアクセス等の外部からの攻撃により、データの書き換えやモデルの変更が行われることで、攻撃者の狙いに応じて、不公正バイアスが引き起こされる恐れがある。

5――不公正バイアスへの対処

AIシステムの運用において、不公正バイアスを完全に排除することは難しい。しかし、適切な対処を行うことで、その発生リスクを軽減することはできる。本章では、軽減策について見ていこう。

1｜規制やルールの変更に対応する

すでに述べたとおり、不公正バイアスの基準は、時とともに変化していく。その変化に応じて、社会で、関連する規制やルールが見直されることもある。そうした規制やルールの変更に対応することで、不公正バイアスの発生を抑えることができる。

2｜スリーラインズを活用する

通常、ERM(全社的リスク管理)では、担当部門、リスク管理・法務・コンプライアンス部門、内部監査・倫理担当部門の“スリーラインズ”によって、重層的にリスク管理が展開されている。不公正バイアスについても、スリーラインズを活用することが不可欠となる。具体的には、次のようになる。

(1) AIモデルの作成・管理担当部署

モデル開発とデータ使用、モデルリスク評価、プロセスの文書化、継続的なモニタリングを実施。

(2) リスク管理・法務・コンプライアンス部門

受入基準とモデル要件を定義し、独立した精査を実施。モデル設計とパフォーマンスの課題を指摘し、課題が解消した場合は承認する。

(3) 内部監査・倫理担当部門

管理とプロセスを精査し、特に機密性の高いAI活用のケースを調査。併せて、保険会社の戦略、目的、ミッションに沿った意思決定を評価。

3｜すべての従業員の倫理学習を充実させる

すでに見たとおり、不公正バイアスは、元をただせば人間に依拠するものと言える。また、直接顧客に対応する場面だけでなく、幅広い業務に潜在している。さまざまな業務でAIの活用が進んでいる現状を踏まえれば、保険会社のすべての従業員の倫理学習を充実させて、偏見や差別に対する認識を高めることが、不公正バイアスの軽減に有効となる[4]。

4｜AI開発では、多様なスキルを持つ従業員の協働を促す

AIを活用したモデル開発を行う際、さまざまなスキルを持つ人材を開発チームに含めることで、複

[4] ただし、倫理学習で従業員の課題の認識向上を図ることはできても、人が持つ偏見を完全に排除できるわけではない点に留意が必要となる。

数の観点から不公正バイアスを評価し、リスクを軽減することができる。例えば、自動車保険で、給付の不正請求の支払査定のトリアージ[5]のための AI モデルを開発する場合、開発チームに自動車保険詐欺を検出するスキルを持っている人材を入れる。こうすることで、システム開発者とは異なる視点から、不公正バイアスの混入を評価できるようになり、リスクの軽減が図られる。

5｜AI の不公正バイアスは、モデルリスク管理の一環として行う

一般に、リスク管理全体の時間・手間・コストを踏まえると、AI の不公正バイアス混入だけを単独で検出・評価して、対応することは、あまり上手なやり方とは言えない。通常、AI のモデルには、モデルリスク管理が行われる。そこで、そのリスク管理の一環として、不公正バイアスのリスクを管理していくことが考えられる。このように既存のリスク管理手法に統合させることで、すでに確立している PDCA サイクルや文書化の手法を活用して、効率的に不公正バイアス混入に対応することができる。

6――おわりに（私見）

以上、AI の不公正バイアスの問題と対処方法について見ていった。この問題は、AI を活用するなかで発生する。このため、システム部門以外の職員は、「AI の不公正バイアス問題は、AI 開発の問題であって、自分には関係ない」ととらえがちになる。しかし、本稿で見てきたとおり、不公正バイアスは、元をただせば人間に依拠するものと言える。AI が機械学習で用いるデータのなかに、不公正バイアスの芽が入り込んでいる可能性がある。

したがって、従業員を対象とした倫理学習を定期的に行い、偏見や差別に対する認識を高めることが、大きなカギとなる。

今後、保険会社のみならず、社会のさまざまな場面で、AI の活用がさらに進んでいくものと考えられる。不公正バイアスの問題も、いま以上に、重要な課題となっていく可能性がある。引き続き、AI の活用と、その課題への対処について、保険会社等の取り組みの動向を注視していくこととしたい。

（参考文献）

“Avoiding Unfair Bias in Insurance Applications of AI Models”（SOA Research Institute, Aug. 2022）

“Fair Work Act 2009”（Section 351(1), Australia）

[5] 優先順位付けをすること。不正請求の可能性が高い事案を優先して、慎重な支払査定を行う。