AIの公平性～差別のないAI社会の実現に向けて～

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）

○　AI の活用が進む中で、AI ガバナンスに注目が集まっており、日本国内においても AI ガバナンス体制の整備を進める企業がみられる。AI ガバナンスの文脈において、AI に差別や偏見がないかを担保したり、評価したりする「AI の公平性」が求められるようになっている。

○　欧米では「AI の公平性」を法律として規制する動きがあり、違反すると巨額の制裁金や業務停止命令を受けるリスクが現実的なものになりつつある。公平性を欠いた AI サービスを提供することで、企業の信頼を失うような事例も欧米ではみられており、今後 AI サービスの活用が進む日本においても同様の事例が起こる可能性がある。

○　AI の公平性をはかる様々な指標を理解し、保有する AI を評価・検証することで、AI の公平性にかかるリスクを低減させていくことが、AI 提供者において、今後特に重要となることが推察される。

1. 「AIの公平性」が求められる背景

「AI の公平性」を端的にいうと、「AI モデルの予測結果に差別や偏見がないかを評価し、是正するもの」である。AI というと、これまではどの程度の精度がでるかという点に着目されることが多かったが、近年では精度のみならず、本レポートのテーマである「AI の公平性」のような観点にも注目が集まっている。そこでまず、「AI の公平性」が求められるようになった背景として、筆者が考える 3 つのポイントを説明したい。

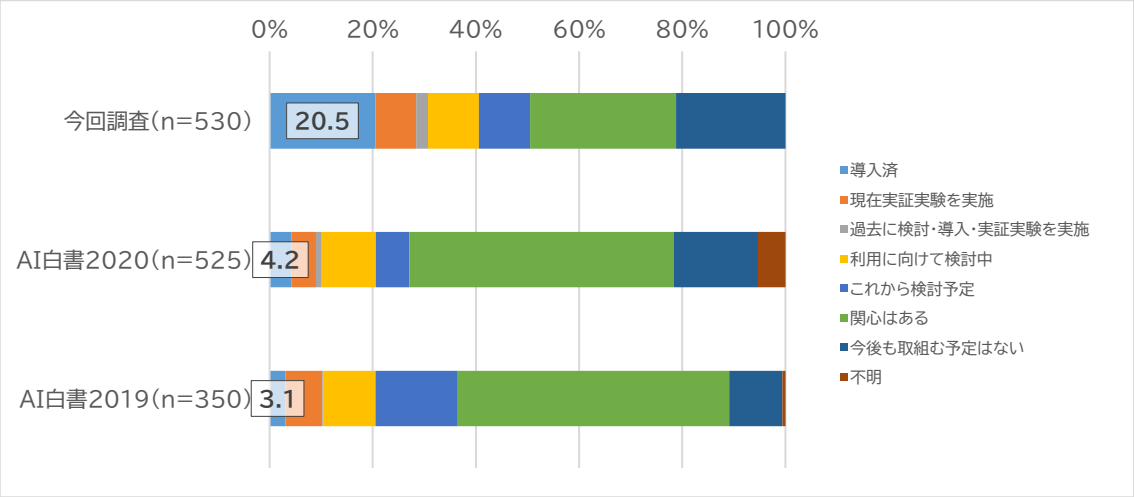
1 点目は、AI が実用フェーズに突入したという点である。これに伴い、AI のガバナンス・倫理・運用といった観点がより一層注目されるようになってきている。

国内においても、内閣府の「人間中心の AI 社会原則」や、総務省の「AI 利活用ガイドライン」などを参考にしながら、AI ガバナンスを検討する企業が増えている。いずれの指針についても、公平性、説明責任、透明性、プライバシーといった観点に配慮しつつ AI を提供すべきと示しており、公平性を含む AI ガバナンスの構築は、企業にとっても重要なキーワードとなっている。

AI の活用が進み、AI を試しに作ってみるフェーズから、実用フェーズに突入しつ

つあることがわかるデータを紹介したい。資料 1 は、日本における AI 技術の活用状況の経年比較である。これによると、AI を導入している日本企業が 2020 年の 4.2%から 2021 年の 20.5%と大幅に上昇している。こうしたデータより、AI の活用が浸透していることがわかるだろう。

資料 1　日本における AI 技術の活用状況



(出所)IPA「DX白書 2021」図表 14-5 のデータを元に筆者作成

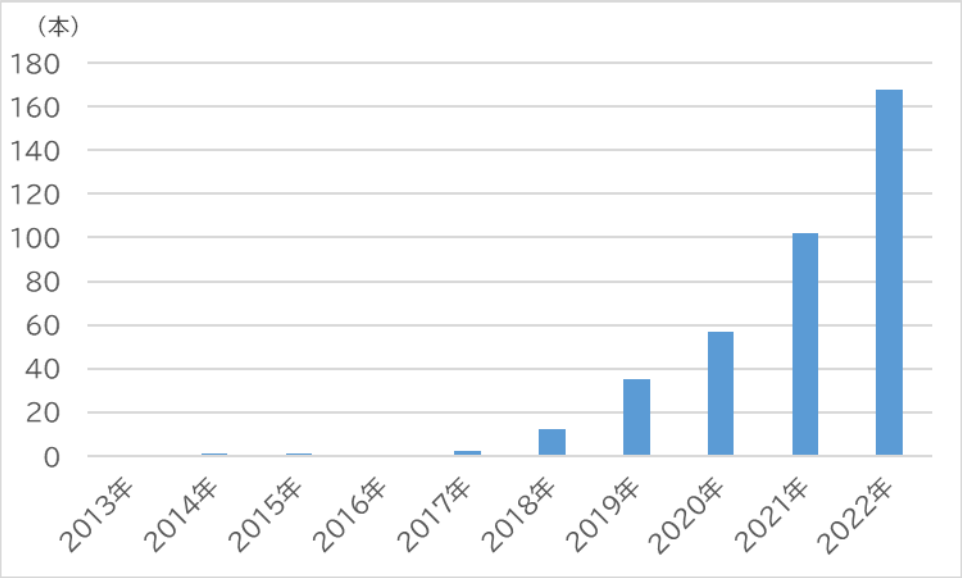
2 点目は「AI の公平性」が問題となった事例が実際に発生している点である。詳細は次節で説明するが、「AI の公平性」を欠いた AI サービスを提供したことで、一般ユーザーからの信用を失う事例が見られ始めている。そして 3 点目は、そうした事例に伴い「AI の公平性」に対する法規制が各国で着々と進んでいる点である。

以上述べたように、AI が実用フェーズに突入し、AI ガバナンスや運用に関する議論が重要視されるようになり、実際の事例や法整備も相まって、「AI の公平性」が求められるようになってきている。

2. 「AI の公平性」の動向

本節では「AI の公平性」が求められる背景を示した上で、実際の動向をみていきたい。まず学術分野の注目度を確認する。資料 2 は、arXiv という論文検索サイトで「Fairness（公平性）」と「AI」が概要に含まれる論文を年ごとに検索した際の件数推移である。右肩上がりで論文数が増えており、注目度が高まっているテーマであることがわかるだろう。

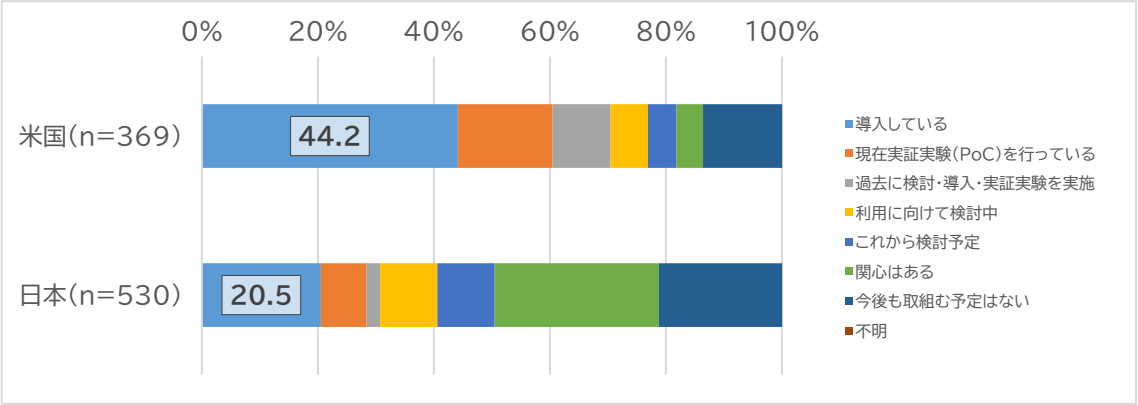
資料 2　論文検索サイト「arXiv」における「AI の公平性」に関する論文数



（出所）論文検索サイト「arXiv」のデータを元に筆者作成

次に事例についても確認していく。前節で述べたように AI のサービス利用が広がる一方、問題も出始めている。例えば、米国の先進的な IT 企業による AI を活用した人材採用システムが女性を差別していたことが発覚し、2017 年にプロジェクトが終了した。また 2019 年に米国の別の先進 IT 企業が提供するクレジットカードの AI を活用した審査において、与信スコアは妻の方が高いにも関わらず、夫が妻より 20 倍高い利用限度額が設定されていたことが、カード利用者の SNS への書き込みによって発覚した事例などが存在する。日本において、レポート執筆時点では「AI の公平性」に直結する事例で目立ったものは存在しないが、ここで考慮したいのは、テクノロジーが浸透していくスピードにおいて、日本は欧米を追従するという点である。資料 3 は日米における AI 技術の活用状況の比較である。AI を導入している企業の割合をみると、米国企業が 44.2%と、日本企業の 20.5%を大きく上回っていることがわかる。今後、日本でも米国並みに AI 活用が進むと、米国で先行したような事例が日本でも起こる可能性は十分にあるといえるだろう。

資料 3　日米の AI 技術の活用状況（2021 年）



（出所）IPA「DX 白書 2021」図表 14-5 のデータを元に筆者作成

続いて各国の法規制という観点でみてみたい。欧米では法規制が進んできており、特に影響が大きいとされるのが 2024 年に全面施行が予定されている EU の「欧州 AI 規則案（EU AI Act）」である。これは、EU 域内の人を対象に AI システム自体や AI を用いたサービスを提供した場合、日本の事業者であっても適用されるという点から、特に注目したい規制である。内容としては、AI をリスクベースで分類（注 1）し、それに応じて規制内容を変えるものになっている。規制対象となった場合、AI の精度や説明性、そして本レポートのテーマである「AI の公平性」についても検証およびモニタリングをすることが求められている。なお違反した場合には巨額の制裁金や業務停止命令が下される可能性があるため、実務面でも緊急性の高い課題であるといえるだろう。

また米国においても、2021 年に米国連邦取引委員会（FTC）が FTC 法で禁止している不公平や詐欺的な行為が AI を用いたサービスも対象になると公式サイト上でアナウンスしていたり、州レベルで公平な AI の活用に向けた独自の法案提出が見られたりと、世界的にも「AI の公平性」が求められている状況である。一方、日本においては、ガイドライン等が出始めているが、実効性のある規制という形式では、レポート執筆時点では整備されていない。

3. 「AI の公平性」とは

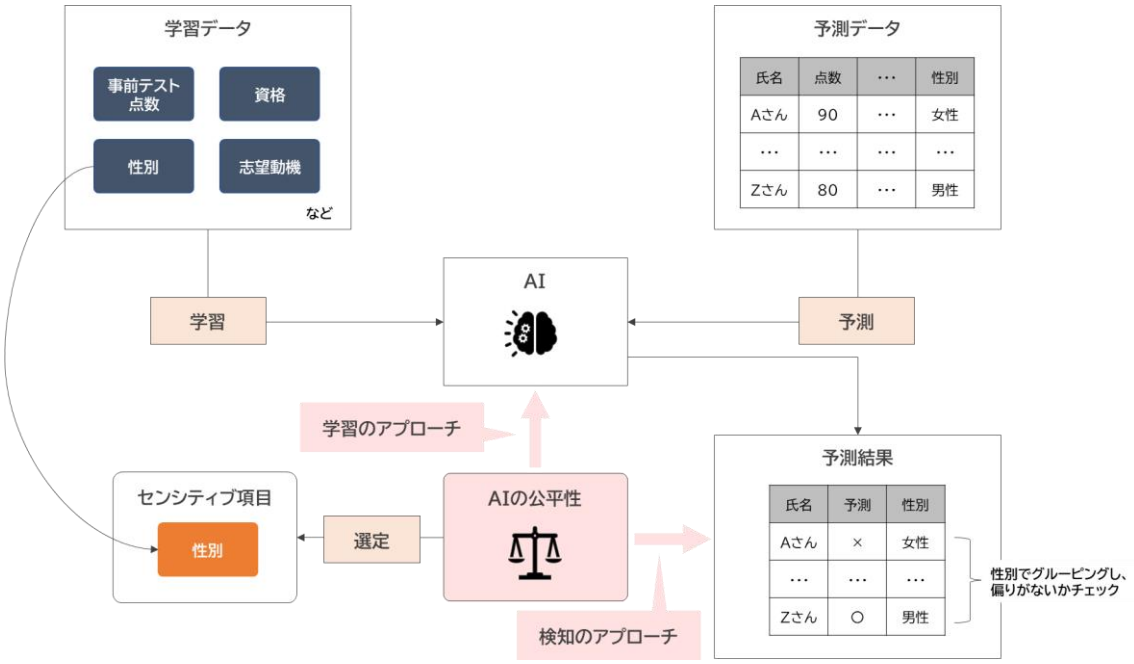
では改めて、「AI の公平性」とは何かを確認していく。冒頭にも述べたように「AI の公平性」を一言でいうと、「AI モデルの予測結果に差別や偏見がないかを評価し、是正するもの」である。AI は膨大な量の学習データやエンジニアが設計したアルゴリズムから構築されるため、それらに偏りがあると、AI の予測結果に差別や偏見が生じてしまう可能性がある。それを見つけ出し、是正するものが、「AI の公平性」である。

そして差別や偏見がないかの評価対象となる項目のことを「センシティブ項目」と呼ぶ。センシティブ項目については、人種・性別・宗教・肌の色などが例として挙げられる。しかし、共通の定義は存在しておらず、AI の利用目的や用途によって、AI サービスの提供者が個別に定める必要がある点は注意が必要である。

この点について、いくつかの AI を例に「性別」をセンシティブ項目とすべきかについて考えてみたい。疾病予測のような AI では、AI の予測結果が性別によって差異があっても、それが生物学的な差異に起因するため、差別にはならないといえるだろう。一方で採用や与信判定の AI において、性別によって AI の予測結果に差異が生じた場合、差別と捉えられる可能性が高い。前者であれば、性別をセンシティブ項目とする必要はないが、後者であればセンシティブ項目とすべきとなる。このように同じ性別であっても差別となるケースと、差別ではない適切な区別となるケースが存在するため、センシティブ項目は AI 提供者が社会的な見方・考え方を踏まえつつ各自で判断する必要がある。

続いて、「AI の公平性」を実現するアプローチを紹介したい。大きく 2 つのアプローチが存在しており、1 つは検知、もう 1 つは学習である。検知については、学習データや AI の予測データ（注 2)、さらには予測が正しかったかどうかのフィードバックデータを用いて、様々な評価指標を用いながら、AI に差別や偏見が含まれるかどうかを検知するというアプローチである。一方、学習については、学習時にデータやアルゴリズムを調整し、差別が含まれない AI を構築しようとするアプローチである。「AI の公平性」の AI モデルに対する位置づけと各アプローチの関係については、資料 4 を参照いただきたい。いずれのアプローチも重要であるが、学習時のアプローチは AI の仕組みに影響を受けるため、今回は AI の仕組みによらず、汎用的に利用できる検知のアプローチについて、いくつかの評価指標の考え方を次節で紹介したい。

資料 4　採用 AI における「AI の公平性」のイメージ



（出所）筆者作成

なお「AI の公平性」の類似の用語として、「AI の説明可能性」がある。「AI の説明可能性」を一言で言うと、「AI モデルやその予測結果において、どの項目がどの程度影響したかを評価・検証するもの」であり、「AI の公平性」とは異なる領域をスコープとしている。前回のレポートでは「AI の説明可能性」を解説しているため、関心がある方は、そちらも参照いただきたい。（AI 活用の鍵「説明可能な AI」とは - https://www.dlri.co.jp/report/ld/185396.html）

4. 「AI の公平性」を実現する評価指標

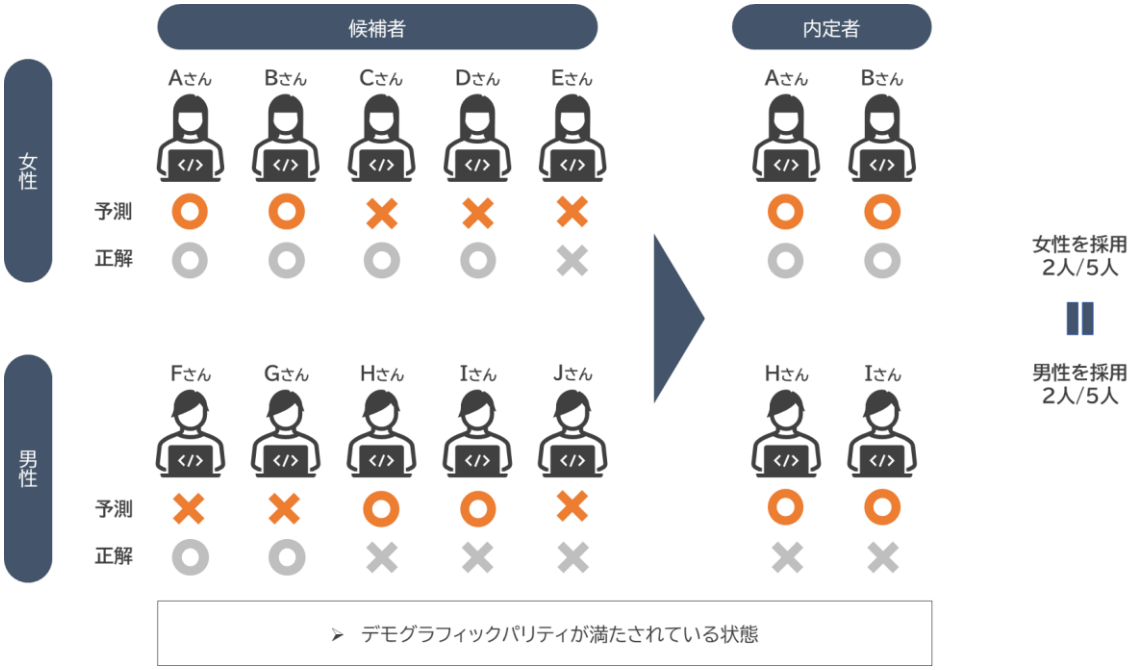
評価指標だけをとっても様々なものが存在するが、今回は特にメジャーなデモグラフィックパリティ、機会均等、等価オッズの 3 つの指標についてみていきたい。説明をするにあたって、採用 AI をユースケースとし、センシティブ項目として性別を選定した。今回の例では男女 5 人ずつの応募者がおり、全体で 4 人の採用枠に対し採用 AI を適用し、性別というセンシティブ項目でグルーピングした上で、グループ単位で公平性をチェックするケースをイメージいただきたい。また図中の正解という項目については、AI が予測する際にはわからない、その人が本来採用すべき人間かどうか（能力的に優れているなど）という正解データを示すものであり、機会均等や等価オッズの評価時に利用する。

なお、今回の例では採用 AI をユースケースとしているが、採用に AI を用いるべきかどうかは別問題であり、あくまでも説明のための例として参照いただきたい。

(1) デモグラフィックパリティ

デモグラフィックパリティを端的に説明すると、「グループごとの予測の比率が同じであるほどよい」という指標である。例示のケースでみていくと、性別ごとの予測の比率が近いほどよいので、男性２名、女性２名を採用すると予測できれば、公平性があると評価される。

資料５　デモグラフィックパリティによる公平性評価イメージ



（出所）筆者作成

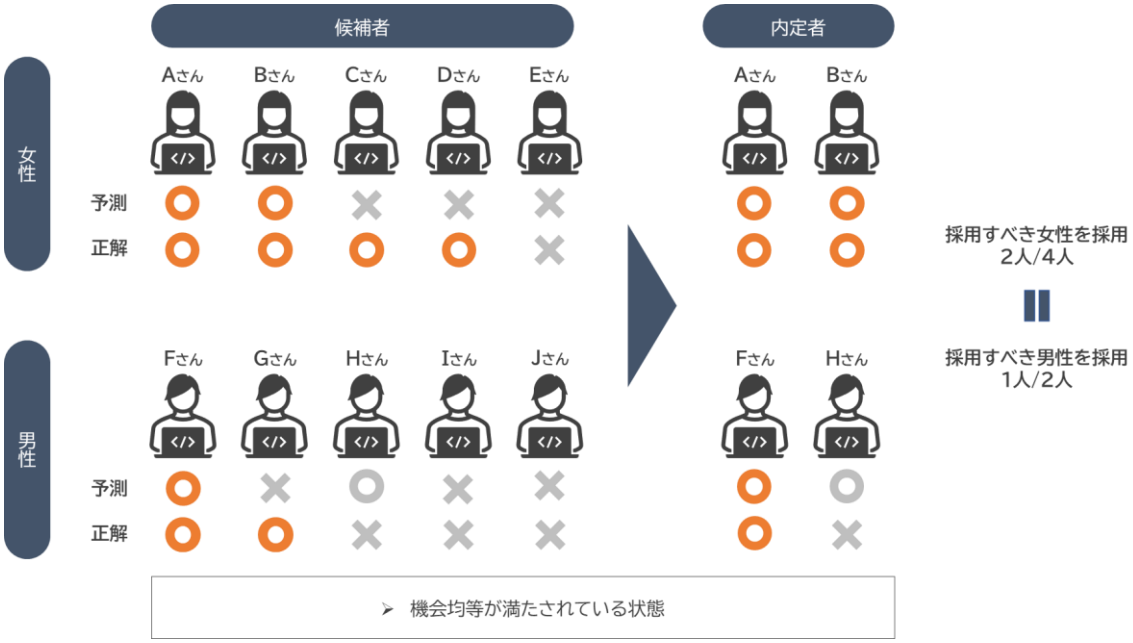
本指標でポイントとなるのは、指標を算出するのにフィードバックデータ（AI の予測が正解だったかどうかの情報）が不要となることである。AI 構築時は正解データ（＝フィードバックデータ）を使って AI モデルを構築することになるため、正解データが手元にあるのが通常である。一方、サービスとして提供すると、AI の予測結果が正しかったかどうかを示す正解データが得られないケースもある。本指標のメリットは予測データのみで評価可能なため、正解データが得られないケースでも公平性のチェックが可能という点にある。

(2) 機会均等

機会均等を端的に説明すると、「グループごとの正解に対して予測の比率が等しいほどよい」という指標である。例示のケースだと、本来採用すべき人（正解）が女性は４名、男性は２名いた場合、AI がその中から女性２名、男性１名を予測した状態

で、男女のグループ間で同じ比率となり、公平性があると評価される。前述のデモグラフィックパリティのケースでは、AI の予測結果のみが評価の対象であるため、極論すると精度を考えずにセンシティブ項目が平等になるように予測するような AI を作ればデモグラフィックパリティを満たすことができた。それに対し、機会均等は、本来権利を有するべき人が、公平に扱われているかどうかを見ることができる指標であるため、精度も意識しつつ公平性も評価できるといえるだろう。

資料 6　機会均等による公平性評価イメージ

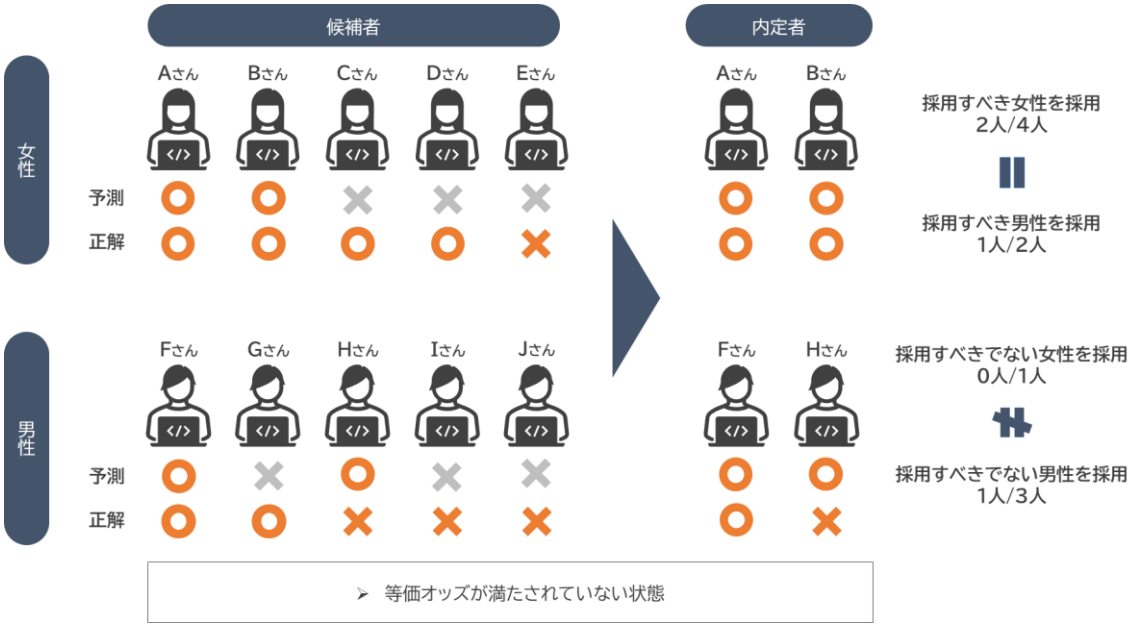


（出所）筆者作成

(3) 等価オッズ

等価オッズを端的に説明すると、「グループごとの正解と不正解に対して予測の比率が等しいほどよい」という指標である。例示のケースに当てはめると、まず本来採用すべき人（正解）の男女の比率については、前述の機会均等で確認したように同じ比率になる。一方、本来採用すべきでない人（不正解）をみると、女性は 1 名、男性は 3 名存在しているのに対し、AI はその中から女性 0 名、男性 1 名を予測しており、男女のグループ間で差異があるため、等価オッズを満たしていないといえる。このように等価オッズは、機会均等に比べて公平性という意味ではより厳格な指標であるといえる。

資料 7　等価オッズによる公平性評価イメージ



（出所）筆者作成

以上、主要な公平性指標について解説した。注意すべきは、全ての AI であらゆる指標をクリアすべきものではない、ということである。一般に公平性と精度はトレードオフと言われており、各指標においても両立しうるものではない。AI の要件・利用用途・目的などに応じて、各指標を選定し、適切に管理していくことが必要である。

本節はやや技術寄りの話になってしまったが、筆者が業務で評価した AI の公平性を可視化するツールにおいても、上記の指標が当然のように使われていた。将来的にはこうした指標を知らないユーザーでもわかるようなツールが登場することが予想されるが、現状においては基礎知識として持ち合わせるべき情報と考え紹介した。

5.　おわりに

「AI の公平性」の評価対象となるセンシティブ項目については、前述したとおり、AI の利用用途や目的に応じて各自で定めなければならない。また公平かどうかを判断するための各指標の基準値にどの程度の値を設定すればよいかの最終判断も AI 提供者に委ねられる。こうした課題については、公平性に影響しそうな項目を自動で抽出し、その中からセンシティブ項目を提案してくれたり、センシティブ項目の内容や AI の中身を考慮して指標値ごとの基準値をレコメンドしてくれるようなツールの登場に期待したい。

加えて「AI の公平性」では、AI の判断が公平かどうかについて、最終的にその AI を利用する一般ユーザーがどう捉えるか、という点も非常に重要になる。そのため

AI サービスの提供者側である企業が定めたセンシティブ項目や基準値と、利用者側である一般ユーザーの感覚が乖離すれば、不公平な AI を提供する企業として、信頼を失う可能性がある点には特に注意が必要であろう。AI サービスの提供者側からすると非常に難しい問題であるが、こうした不確実なものを扱っていくうえで、AI のポリシーを明確にして公開するというアプローチも提唱されているので参考にしたい。例えば AI 機能を搭載したチャットボットにおいて、「学習中のため不適切な回答がありましたらご連絡ください」といった文言があるだけで、だいぶ印象が変わるだろう。また例示したような採用 AI において、例えば現在の男女社員比率に不均衡があり、それを改善するために採用 AI を導入し、男女で採用比率が異なっている場合、その採用方針を明記することで、受け手であるサービス利用者（＝就職希望者や男女の採用比率が異なることを知った一般ユーザー）側のイメージは別物になるだろう。

このように、実際に「AI の公平性」をサービスとして提供する AI の中に取り込んでいくことにはいくつかハードルがあるため、事例やツールが出そろってから検討をすればよいと感じた方もいるかもしれない。しかし「AI の公平性」は、不公平な AI を提供したという問題が発覚してから対応するのでは遅く、また信頼回復のための労力も相当なものになる。これから AI サービスの提供を検討している方は、「AI の公平性」の内容を理解し、提供しようとしている AI が公平かどうか、まず確認してみることの意義は大きいといえるだろう。

以　上

【注釈】

1) 欧州 AI 規則案（EU AI Act）では、AI をリスクの大きさに応じて、「許容できないリスクのある AI」、「ハイリスクのある AI」、「限定リスクのある AI」、「最小リスクの AI」の 4 つに分類している。「許容できないリスクのある AI」は、公的機関のソーシャルスコアリングや、利用者の年齢や障害という脆弱性を利用した AI などが該当し、利用を禁止している。「ハイリスクのある AI」は、インフラや教育、雇用などを対象とした AI が該当し、規制対象となる。「限定リスクのある AI」は、チャットボットのような自然人と相互作用する AI などが該当し、AI であることを通知する義務など、透明性の義務が課される。これらに該当しない AI は、「最小リスクの AI」に分類され、規則案上の法的義務は課されないが、行動規範の奨励というかたちで、ハイリスク AI に求められるガバナンス要件を任意に適用することを促している。
2) AI の出力結果のこと。AI の出力は学習データをもとにした予測であり、その予測を採用するかどうかは、人や連携するシステムが判断する。

【参考文献】

- 独立行政法人 情報処理推進機構(IPA) DX 白書 2021 (2021)
  https://www.ipa.go.jp/publish/wp-dx/dx-2021.html
- 内閣府 統合イノベーション戦略推進会議「人間中心の AI 社会原則」(2019)
  https://www8.cao.go.jp/cstp/aigensoku.pdf
- 総務省 AI ネットワーク社会推進会議「AI 利活用ガイドライン」(2019)
  https://www.soumu.go.jp/main_content/000637097.pdf
- TDSE マガジン - [入門]公平性 (Fairness) とは (2020)
  https://www.tdse.jp/blog/tech/3273/
- 経済産業省「EU の AI に関するフレームワーク」(2021)
  https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/2021_001_05_00.pdf
- 総務省「EU の AI 規制法案の概要」(2022)
  https://www.soumu.go.jp/main_content/000826707.pdf
- 機械学習と公平性シンポジウムについて - Preferred Networks Research & Development (2020)
  https://tech.preferred.jp/ja/blog/ml_and_fairness/
- 【JSAI2022】AI 時代に考えるべき公平性とは何か？ | モリカトロン AI ラボ(2022)
  https://morikatron.ai/2022/08/jsai2022_ml_and_fairness/