機械学習による予測・推論の公平性：金融サービスにおいて求められる配慮とは

宇根正志（うねまさし）

要　旨

本稿では、機械学習の予測・推論を融資判定のサービスに利用するケースを例に、予測・推論の偏りが融資判定の公平性に及ぼしうる影響（典型的には人種データの扱いに伴う問題）、技術的な対応方法、残されている課題を解説する。まず、融資判定の文脈での公平性に関するサービス要件を、原則・社会規範、サービス利用者等の期待に基づいて決定する方法を示す。次に、決定したサービス要件を前提に、公平性に配慮した機械学習利用システムの開発・運用方法や予測・推論の偏りを軽減する各種手法を、機械学習品質マネジメントガイドライン等を参照しつつ解説する。最後に、技術的対応だけでは解決が難しい問題として、公平性に関するサービス要件の設定やそれを具体化する概念の選択に焦点を当てて、サービス利用者との対話を通じて公平性のサービス要件を抽出する手法の研究等、最新の研究成果の概要と課題を解説する。

キーワード：　ガバナンス・ガイドライン、機械学習、機械学習品質マネジメントガイドライン、公平性、融資判定

本稿の作成に当たっては、小西弘一氏（国立研究開発法人産業技術総合研究所）、中尾悠里氏（富士通株式会社）から貴重なコメントを頂戴した。記してここに感謝する。ただし、本稿に示されている意見は、筆者個人に属し、日本銀行の公式見解を示すものではない。また、ありうべき誤りはすべて筆者個人に属する。

宇根正志　日本銀行金融研究所参事役（E-mail: masashi.une@boj.or.jp）

1. はじめに

近年、金融サービスにおける既存の事務の自動化や新しいサービスの創成・提供を企図して機械学習を活用する動きが広がっている。例えば、自動応答（チャットボット）、コールセンターにおける応答対応、マーケティングにおける取引情報の分析、信用度評価、融資判定等が挙げられる（井上・宇根［2020］）。

機械学習は、人間にとって難しいと考えられている問題の解の候補を効率的に提示してくれる。もっとも、Rea [2020] や Agarwal and Mishra [2021] で指摘されているように、機械学習の予測・推論において意図しない偏りが生じうるというマイナス面も存在している。例えば、米国の住宅ローンにおけるデフォルト率を、高度な機械学習の手法（ランダム・フォレスト等）によって予測した研究（Fuster *et al.* [2022]）が知られている。Fuster *et al.* [2022] は、公平性・公正性の観点から使用に配慮が必要な属性である借入申込者の人種のデータを訓練データとして用いていなかったにもかかわらず、デフォルト率の予測値が借入申込者の人種によって有意に偏っていたとの結果を発表している。

こうした研究を踏まえると、サービス利用者への商品・サービスの提案や信用度の評価等、サービス利用者の属性に応じて内容・結果が変化する業務やサービスに機械学習を適用する際には、予測・推論に意図しない偏りが生じる可能性に留意する必要があるといえる。特に、そうした偏りが、人種、性別のように配慮が必要な属性に関係するものである場合には、そのままサービス内容に反映されてしまうと、サービス利用者への差別につながるおそれがある。ただし、サービスの質向上、例えば、リスクに応じた適切な融資判定や金利設定において、配慮が必要な属性が明らかに有益な情報をもたらしている場合、公平性と効率性の間にトレードオフ関係が生じる。これは、「法と経済学」という学問領域で研究されてきたテーマであるが、本稿では議論に立ち入らず公平性の観点からの分析を取り上げる[1]。

機械学習を適用するか否かにかかわらず、金融サービスを提供する際に公平性に配慮することは、「金融サービス業におけるプリンシプルについて」（金融庁［2008］）が示すように、従来から金融機関に求められる規範である。また、最近では、「人間中心の AI 社会原則」（以下、AI 社会原則。統合イノベーション戦略推進会議［2019］）が発表され、公平性を確保することが、AI の研究開発や社会実装における基本原則の 1 つとして掲げられている。金融サービスを含む、各種サービス一般において公平性への配慮が求められていると考えられる。

1 公正性（公平性）と効率性は中学校の公民分野の教科書・指導要領でも取り上げられている重要な論点である。適正な金融サービスの在り方を考える際には、別途、考察が必要な分野である。公平性と公正性の概念については後述する。

では、公平性に配慮した機械学習のシステム開発やサービス提供をどのように行うことができるのであろうか。従来のシステム開発・運用のプロセスを前提にすれば、まず、公平性への配慮として求められる事項をサービス要件として設定する。次に、サービス要件を満たすうえで、機械学習を実装するシステムに要求される事項を非機能要件として設定し、それを満たすようにシステムを開発・更新してサービスを提供するという流れが考えられる。

ただし、次の 2 点が課題となる。1 つは、公平性のサービス要件をどう設定するかである。公平性への要請の内容が法律や規範として明文化されている場合、それを遵守することが求められるが、それだけでは必ずしも十分とはいえない。一般に、公平性の捉え方は、サービス利用者をはじめとする関係者（ステークホルダー）によって異なるほか、個人によっても、国や地域社会によっても、時代によっても異なりうる。個々の金融サービスの事例に応じて、各ステークホルダーが望む公平性への配慮を適切に把握し、サービス要件化することが必要である。

もう 1 つは、公平性のサービス要件を特定できたとして、それを満たすためにはシステムをどのように開発・運用すればよいかである。機械学習を実装するシステムの開発・運用に関して、最近、複数のガイドラインが策定・公開されている。機械学習品質マネジメントガイドライン（以下、品質ガイドライン。国立研究開発法人産業技術総合研究所［2022］）、AI 原則実践のためのガバナンス・ガイドライン（以下、ガバナンス・ガイドライン。AI 原則の実践の在り方に関する検討会［2022］）、AI プロダクト品質保証ガイドライン（AI プロダクト品質保証コンソーシアム［2023］）である。これらはいずれも汎用性が高く、金融分野にも適用可能であるものの、実際に活用する際には記載事項（推奨事項等）を個々の金融サービスの文脈で解釈する必要がある。

本稿では、上記の課題への対応について、金融サービスの典型的な事例として融資判定に機械学習を活用するケースを想定して解説する。2 節では、機械学習を実装したシステムの構成や開発・運用プロセスを説明する。3 節では、機械学習が融資判定の公平性に与えうる影響や、公平性に関するサービス要件の決定方法を説明する。4 節では、ガバナンス・ガイドラインや品質ガイドラインを参照しつつ、システムの開発・運用上、公平性にどのように配慮すればよいかを解説する。5 節では、3、4 節を振り返りつつ、2 つの課題への対応に関して考察して本稿を締め括る。

なお、公平性と類似した概念として公正性が存在する。公平性は機会や結果の公平さを指す帰結主義的な考えに沿うものであり、どのような帰結が善いものであるかという判断を内包する。一方、公正性については、公平になるよう正しく扱うという手続き主義的な考え方の意味合いが強まる[2]。こちらも手続きの正しさとは何かという論点を内包している。機械学習における公平性・公正性を巡る日本語の議

2 公平性には、結果に注目した分配的公正性と、手続きに注目した手続き的公正性がある。

論では公平という用語が使われることが多いため、本稿ではそうした用語法に従って公平性を用いる。ただし、手続き主義的な考え方を指している文脈では公正性を用いる。

2. 機械学習利用システム

本節では、機械学習（ここでは教師あり学習）を実装したシステムの基本的な構成と開発・運用プロセスを説明する。以下では、機械学習を実装したシステムを、品質ガイドラインの用語に倣って、機械学習利用システム（machine learning based systems）と呼ぶほか、その他の構成要素についても品質ガイドラインの用語に基づいて説明する。

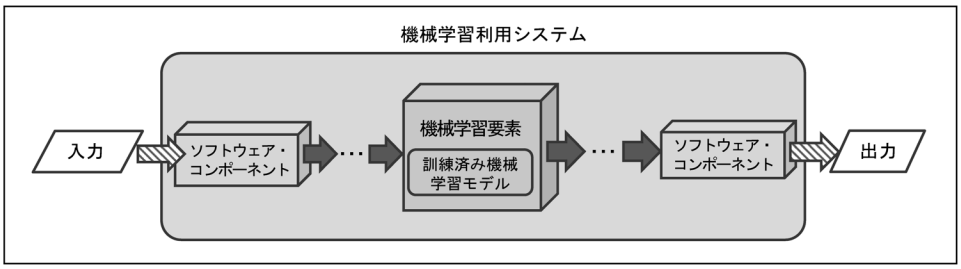
(1) 基本的な構成

機械学習利用システムは、外部から入力を得て、それに対する予測・推論の結果を出力するシステム[3]である（図1を参照）。具体的には、①外部から入力を受け取るソフトウェア・コンポーネント、②機械学習技術を応用し、入力から予測・推論結果を生成するソフトウェア・コンポーネント（機械学習要素〈machine learning component〉）、③機械学習要素の出力を処理して外部に予測・推論結果を出力するソフトウェア・コンポーネントが含まれる。

機械学習利用システムのソフトウェア・コンポーネントのうち、予測・推論を行うという点で、機械学習要素が重要な役割を担う。機械学習要素は、予測・推論の動作を規定する各種パラメータ（訓練済み機械学習モデル〈trained model〉）をソフトウェアとして実装したものである。訓練済み機械学習モデルは、訓練データを用いて生成される。

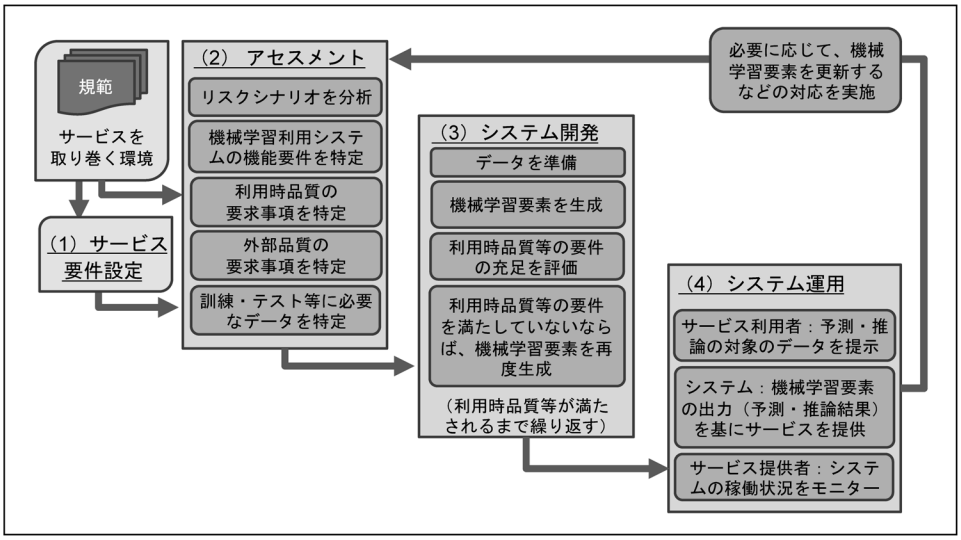
3 機械学習利用システムには、後述する機械学習要素による予測・推論を主な目的とするものだけではなく、その予測・推論の結果を用いて別の処理を実行することを主な目的とするものもある。例えば、機械学習を用いた自動運転システムでは、外部の環境の情報（画像等）を入力として取得し、取得した情報を機械学習要素によって認識しつつ先行きを予測したうえで、（主たる目的である）運転にかかる処理を実行する。本稿では、金融サービスに適用するケースを取り上げることから、予測・推論を主な目的とする機械学習利用システムに焦点を当てる。

図 1　機械学習利用システムの基本的な構成（概念図）



資料：国立研究開発法人産業技術総合研究所［2022］

図 2　機械学習利用システムの開発・運用のプロセス（イメージ）



資料：宇根・清藤［2020］

(2)　開発・運用

機械学習利用システムの開発や運用について、丸山［2017］や宇根・清藤［2020］は、サービス要件を設定したうえで、アセスメント、システム開発、システム運用の各フェーズを実施するケースを紹介している（図 2 を参照）。公平性に関する問題を検討する前に、一般論として各フェーズの内容を紹介し、3 節以降で公平性問題を検討する基礎とする。

イ．サービス要件設定

ここでのサービス要件は、機械学習利用システムの適用対象とするサービスの要求事項であり、サービスを提供する主体（サービス提供者）が決定する。例えば、サービスにおけるビジネス上の目標や、遵守すべき事項（関係する法律や社会規範）等、多岐にわたる事項がサービス要件の候補となる。サービス要件の設定に際しては、サービスの内容に加えて、それを取り巻く環境、想定されるリスク等を考慮することが必要となる。

ロ．アセスメント

このフェーズでは、サービス提供者とシステム開発者（サービス提供者の依頼を受けて機械学習利用システムを開発する主体）が、以下を実行する。

① 機械学習利用システムにおいてリスクとなる事象とそれが顕在化するシナリオを分析する。
② サービス要件と上記①の分析結果に基づき、機械学習利用システムが果たすべき機能の要求事項（機能要件）を特定する。
③ 機械学習利用システムが満たすべきセキュリティや安全性、公平性、有用性（AI パフォーマンス）等の品質（利用時品質と呼ばれる）の要求事項を特定する。
④ 利用時品質要件に基づいて、機械学習要素の出力に関する品質（外部品質と呼ばれる）の要求事項を特定する。
⑤ 利用時品質および外部品質の要件を満たすために必要なデータ（訓練やテストに使用）を特定する。

ハ．システム開発

このフェーズでは、サービス提供者やシステム開発者が以下を実行する。

① 訓練やテストに用いるデータを準備する。
② 訓練データを用いて機械学習要素を生成する。
③ テスト・データを用いて、生成した機械学習要素が利用時品質や外部品質の要件を満たしているか否かを評価する。
④ 利用時品質や外部品質の要件を満たしていないならば、訓練データや機械学習要素の生成方法を再考して変更し、変更後の方法に沿って、機械学習要素を再度生成する。場合によっては、利用時品質の要件等を見直すことも検討する。
⑤ 上記①～④を、利用時品質等が満たされるまで繰り返す。

ニ．システム運用

このフェーズでは、サービス提供者が以下の流れで機械学習利用システムを運用する。

① サービス利用者は、予測・推論の対象となるデータをシステムに提示する。
② 機械学習利用システムは、それを機械学習要素に入力し、機械学習要素の出力（予測・推論結果）を基にサービスを提供する。
③ サービス提供者は、システムの稼働状況（予測・推論結果の内容を含む）をモニターする。
④ サービス提供者は、サービス提供上の問題を検知した際には、必要に応じてアセスメント・フェーズ（の一部）の作業を行い、機械学習要素を更新するといった対応を実施する。

3. 融資判定への機械学習の活用と公平性

本節以降は、金融サービスにおいて機械学習を活用する例として、融資判定に絞って議論を進める。まず、本節では、機械学習を用いた融資判定と、機械学習を用いることによって生じうる公平性への影響、公平性に関するサービス要件の特定の方法をそれぞれ説明する。

(1) 機械学習による融資判定

金融機関が融資判定用の機械学習利用システムを開発するケースを想定する。金融機関は自社が有するさまざまな個人（顧客）の取引や属性に関する情報を訓練データとして用いて訓練済み機械学習モデルを生成する。資金を借りたい個人（借入申込者）は、借入申込時に、自分の属性、借入希望金額やその用途等を金融機関に提示する。金融機関は、それらの情報やその個人の取引情報（例えば、借入残高、返済履歴）を機械学習利用システムに入力し、出力される判定結果（融資可能／融資不可能）を基に最終的に融資の可否を決定する。ここで、機械学習によって融資可否をどのように判定するかは各社のノウハウであり、具体的な方法は通常開示されない。

機械学習を活用した融資判定では、一見、融資判定との直接の関係が薄いようにみえるデータを用いるケースが近年注目されている。例えば、Hurley and Adebayo [2016] や井上・宇根［2020］は、個人のソーシャル・メディアのデータ（例えば、

ソーシャル・メディア上で交流がある個人の属性)、ウェブサイトの閲覧履歴、ウェブ上で購入した商品・サービス、訪問した場所（行動範囲）の情報等を挙げている。こうしたデータが個人の信用力と相関を有しているならば、機械学習がそうした関係を抽出・強調し、融資判定の手掛りを与えてくれる。

新しいデータを活用した機械学習による融資判定は、従来は金融機関と取引関係が薄かった個人も借入のサービスを受けるチャンスにつながるほか、金融機関にとっては、融資対象となる顧客の裾野を拡げる効果が期待される。また、従来よりも多彩なデータと高度な機械学習のアルゴリズムによって、融資判定をより精緻に行うことが可能となれば、貸倒れのリスクの低下も期待されるし、借り手にとってもより低い金利の適用機会が広がりうる。

(2) 公平性への影響～機械学習要素による相関関係の強調

機械学習を融資判定に適用する場合、訓練データと判定結果との間の何らかのパターン性を発見し、これを判定に利用するよう機械学習要素が生成される。そのため、訓練データに含まれる特定の属性が判定結果により強く影響を及ぼすようになる。融資判定自体の研究ではないが、関連研究として Fuster *et al.* [2022] を挙げることができる。Fuster *et al.* [2022] は、住宅ローンのデフォルト率の（機械学習による）予測値が借入申込者の人種によって強く影響を受けるという現象を発見しており、ランダム・フォレストを用いた場合の方が、比較的単純な手法とされていた回帰モデルの場合よりも影響の度合いが大きかった旨を報告している。

Fuster *et al.* [2022] は、機械学習には訓練データと予測・推論結果の間に存在する微小な傾向や偏りを抽出・強調する特性があるとしたうえで、より高い予測・推論精度をもたらす手法ほど、こうした特性が高まると指摘している。同様の見解を Bolukbasi *et al.* [2016] や Zhang, Lemoine, and Mitchell [2018] も示している。

こうした知見を踏まえると、次のような公平性上の問題が懸念される。

① サービス提供者は、融資判定用の機械学習要素を生成するための訓練データを準備した。

② 訓練データには、過去の借入申込の情報、例えば、借入申込者の属性情報や融資判定結果（融資可能／不可能）が含まれていた。

③ 上記の借入申込者の属性情報の一部（属性 X と呼ぶ）は、配慮が必要な属性（性別、国籍、人種等）であって、属性 X の値と融資判定結果が人間では気づきにくい相関を有していた。
—— 例えば、属性 X が国籍、その値として国籍 A と国籍 B が含まれていたとして、国籍 A の場合に融資可能と判定される確率が、国籍 B の場合に融資可能と判定される確率に比べて大きかった。

④ 属性 X を含む訓練データを用いた訓練の結果、属性 X と融資判定結果との間の相関関係を強調するように機械学習要素が生成された。

⑤ 後日、国籍 B の個人が借入申込を行うために自分の属性情報を機械学習要素に入力したところ、融資不可能と判定された。そこで、国籍 A と国籍 B の個人の融資判定結果を基に、融資可能と判定される確率を算出・比較したところ、国籍 A の個人が国籍 B の個人よりも有意に高い確率（確率の差分が 0.3）で融資可能と判定されていたことがわかった。

上記の状況では、属性（国籍）が融資判定結果に有意な影響を及ぼしているが、これが、融資判定が不公平であるということには必ずしもならない。借入申込者やサービス提供者が、この機械学習利用システムおよびそれを利用したサービスに対し、どのような公平性を要求しているかに依存する。仮に、機械学習要素に入力される国籍の値が変わっても、融資可能と判定される確率は有意に変化しないという状況が要求されており（利用時品質要件）、確率の差分が 0.3 であることは許容できないと借入申込者が考えるのであれば、公平性が達成されていないことになる。逆に、確率の差分が 0.3 であるというのは許容できると考えるのであれば、公平性は達成されているとの判断となる。ただし、こうした判断についてステークホルダーの同意が得られるかは別の問題として存在する。

上記の状況において公平性が達成されていないと認識された場合、配慮を必要とする属性 X を訓練データから除去し、改めて機械学習要素を生成するという対応が考えられる。もっとも、これは必ずしも十分とはいえない。例えば、品質ガイドラインは、「実世界から取得したデータは複雑な構造や相関を持ち、直接的な要配慮属性を入力に含めない場合でも、他の入力データから構築した AI が結果的に要配慮属性に相当する値と相関を見いだすことが十分に考えられる[4]」として、「要配慮データを取り扱わなければ自動的に公平になるという立場は取らない」としてい

4 このような効果を、Calders and Verwer [2010] はレッド・ライニング効果（red lining effect）と呼んでいる。

る。このように、配慮が必要な属性以外の属性にも気を配ることが求められる。

(3) 公平性に関するサービス要件

機械学習を活用した融資判定を検討するうえで、借入申込者へのサービスの公平な提供という観点からは、公平性とは具体的にどのような状態を意味するかをまず決める（公平性に関するサービス要件を設定する）必要がある。具体的には、法律や原則・社会規範に基づいて設定するケースと、ステークホルダーの期待に基づいて設定するケースが想定される。

イ. 原則・社会規範に基づくサービス要件

金融機関の融資に関する法律や原則・社会規範が存在し、そのなかで融資判定の公平性に関する規定が存在するならば、金融機関は、法令遵守というコンプライアンスと社会規範逸脱に対する企業評判の両面から、融資判定の方法やそのための情報を検証し開発時のサービス要件に取り込む必要がある。本稿執筆時点では、わが国には、そうした規定が存在していない。一方、原則・社会規範としては、金融サービス業におけるプリンシプルと AI 社会原則が存在することから、これらを参考にすることができる。

金融サービス業におけるプリンシプルは、金融サービスの提供に際し、「利用者の合理的な期待に応えるよう必要な注意を払い、誠実かつ職業的な注意深さをもって業務を行う」ことを要請し、「利用者の公平（な）取扱い」にも触れている。ただし、何をもって「利用者の公平（な）取扱い」とするかに関する記載はない。

AI 社会原則は、公平性の原則として次の 2 点を示している（補論 1 を参照）。

> （A）「AI の設計思想の下において、人々がその人種、性別、国籍、年齢、政治的信念、宗教等の多様なバックグラウンドを理由に不当に差別をされることなく、全ての人々が公平に扱われなければならない。」
>
> （B）「上記の観点を担保し、AI を安心して社会で利活用するため、AI とそれを支えるデータないしアルゴリズムの信頼性（Trust）を確保する仕組みが構築されなければならない。」

（A）は配慮が必要な属性の事例を示している。**こうした属性が融資の判定に影響を及ぼさないようにすること**をサービス要件とすることが考えられる。ただし、

例示されている属性を機械学習に用いないとしても、それと有意な相関を有する別の属性を用いていた場合、例示されている属性が融資判定に影響を及ぼしているようにみえる可能性（レッド・ライニング効果）もある。配慮が必要な属性と関係性を有する属性も配慮が必要な属性に含めるか否かを検討する必要があるが、これは容易ではない。例えば、人種と所得が相関していた場合、所得を配慮が必要な属性とみなして訓練データから排除すると重要な情報を取りこぼす可能性がある。ケースによっては多くの変数が利用できなくなり、機械学習の精度が劣化することも考えられる。これは、1 節で触れた効率性と公平性のトレードオフ問題である。

次に、（B）に対応するために、**配慮が必要な属性が融資判定に影響を及ぼさないように制御していることをステークホルダーに示して理解を得ること**を要件とすることが考えられる。こうしたアプローチは、法哲学の分野では手続き的公正性と呼ばれている。機会や結果の公平性に注目した（A）とは異なる概念である。ステークホルダーの理解を得るためには、公平性への配慮に相当するシステム上の制御やサービス運用上の対応をステークホルダーにある程度開示することが必要となる。さらに、金融機関が、公平性に配慮する取組みの効果を自ら評価し、評価結果を公平性の達成度合いとしてステークホルダーに開示することも有用である。

こうしたアプローチにも弱点が存在する。配慮が必要な属性を用いてはいないが、意図的にそれらと相関している情報、かつ配慮が必要でないと認識されている情報を用いることで、融資判定の精度を上げるという偽装が可能である。公平さの偽装という意味で fairwashing と呼ばれている（荒井ほか［2020］）。こうした偽装を行っていないことを示すためには、手続き的公正性を技術面まで踏み込んだ技術的適正手続き（technological due process）が必要となる。技術的適正手続きでは、アルゴリズムやデータの透明性や精度、説明責任等が求められるため、開示の範囲がより広いものとなる。ただし、疑義を訴えることがコスト負担や不利益の発生の証明義務もなく可能であるとともに、説明責任を開発・利用企業側がすべて負うことになるとすれば、サービス提供が阻害され、利用者便益も実現しなくなるという問題も生じうる。

ロ．ステークホルダーの期待に基づくサービス要件

金融サービス業におけるプリンシプルは、サービス利用者の合理的な期待に応えるよう必要な注意を払うことを金融機関に求めている。ステークホルダーが期待する公平性への配慮も、サービス利用者の合理的な期待に含まれると考えられる。以下では、ステークホルダーの期待をどのようにして把握するかについてガバナンス・ガイドラインを基に説明する。

（イ） 公平性を損なう事案の回避

ガバナンス・ガイドラインは、サービス提供者やシステム開発者が AI システム

に求められる公平性を把握しているかを評価項目例の 1 つとして示し、具体的な確認項目例として以下の主旨の項目を挙げている（補論 2 を参照）。

- 公平性に関するインシデント事例を調査したか。
- サービス対象地域、また、類似のシステムにおいて、偏見や差別的な扱いが生じたといった指摘（の有無）を確認したか。

これらを踏まえると、融資判定やそれに類似するサービス（例えば、クレジット・スコアリング）において、公平性に関するこれまでのインシデント事例や指摘を調査することがまず挙げられる。そうした事案等を見つけた場合には、**融資判定において既知のインシデント（〇〇〇の事例）を回避すること**をサービス要件に追加することが考えられる。

ただし、問題化した事例しか確認できず、公平性を巡る社会規範を悉皆的に洗い出すことはできない。また、どのような事例が問題となるかは時代や環境とともに変化していく。上述の調査や確認には限界があることを認識したうえで活用することになろう。

（ロ） 公平性の捉え方に関するステークホルダーからの聴取

Stumpf *et al.* [2021] は次のようなアプローチを提案している（5 節で説明）。システム開発者が、借り手や貸し手といったステークホルダーに対して、融資判定の機械学習利用システムとそれを用いたサービスの目的、仕組み、制約条件、公平性の観点で配慮する事項等を説明する。そのうえで、ステークホルダーに公平性の観点から意見を求め、ステークホルダーが期待する公平性の配慮を明らかにしてサービス要件とする。

ハ. 公平性に関するサービス要件の例

上記の原則・社会規範やステークホルダーの期待に基づく場合、公平性に関して以下のサービス要件を設定することが考えられる。

（A） 配慮が必要な属性が融資の判定に影響を及ぼさないようにすること。

（B） 配慮が必要な属性が融資の判定に影響を及ぼさないように制御していることをステークホルダーに示し理解を得ること。

（C） 融資判定において、既知のインシデント事例の発生を回避すること。

（D） ステークホルダーが期待する公平性への配慮を実現すること。

4. システム開発・運用プロセスにおける公平性への配慮

本節では、融資判定の文脈において公平性に関するサービス要件を設定済みであるという前提のもとで、機械学習利用システムの開発・運用を進めるうえで公平性への配慮として要求される事項や対応方法を、品質ガイドラインやガバナンス・ガイドラインを参照しつつ各フェーズに沿って解説する。

(1) アセスメント・フェーズ

このフェーズでは、公平性を機械学習利用システムにおける品質と位置付け、品質の要求事項を設定するとともに、訓練やテストに用いるデータを選定する。品質ガイドラインは次の要求事項[5]を示している。

- 機械学習利用システムと機械学習要素における公平性を品質（利用時品質、外部品質）と捉え、それぞれの要件を特定する。
- 訓練データとして採用するデータを検討した後、それに含まれる属性間の依存性や因果関係を明確にする。
- 機械学習要素の出力や訓練データについて、公平性に関する評価指標（公

5 品質ガイドラインは、公平性を確保するための要求事項を 3 つのレベル（AIFL 0、1、2）に分けて記載している。金融サービスの場合、最上位のレベル（AIFL 2）が当てはまると考えられることから、ここでは同レベルに焦点を当てて説明する。詳細は補論 3 を参照されたい。

平性メトリクス）をそれぞれ設定する。

イ．利用時品質等の要件の特定

サービス提供者やシステム開発者は、サービス要件から、公平性の利用時品質と外部品質の要件をそれぞれ導出する。利用時品質とは機械学習利用システムの出力に関する品質を指し、外部品質とは（機械学習利用システム内部に存在する）機械学習要素の出力に関する品質を指す[6]。ここでは、議論を単純化するために、機械学習要素が融資可能あるいは不可能を出力し、これを機械学習利用システムがそのまま出力するケース（利用時品質と外部品質が一致）を想定したうえで、利用時品質要件に焦点を当てて説明する。

融資判定のサービス要件の一部が、**借入申込者の国籍（配慮が必要な属性）が融資の判定に影響を与えないようにすること**であったとする。この「影響を与えない」という部分に関してさまざまな解釈が可能であり、機械学習利用システムの設計に適用できるように具体化する必要がある。これまでに公平性の概念としてさまざまなものが提案されている（補論 4 を参照）。代表的な概念を 4 つ取り上げて利用時品質要件として記述すると、例えば以下のとおりである。

- 例 1（demographic parity）：国籍 A の借入申込者が融資可能と判定される確率と、国籍 B の借入申込者が融資可能と判定される確率に関して、これらの確率の差分（絶対値）が一定値以下となるようにすること。
  —— これは、国籍の値が異なっていても、その他の属性の値が同一ならば、望ましい判定（融資可能）を出力する確率を同一にすべきであるという考え方に基づく要件例である。

- 例 2（equalized odds）：返済能力があり融資対象とすべき個人が融資可能と（正しく）判定される確率を X とし、返済能力がなく融資対象とすべきでない個人が融資可能と誤って判定される確率を Y とする。このとき、国籍 A の個人と国籍 B の個人における確率 X の差分および確率 Y の差分が、いずれも一定値以下となるようにすること。
  —— 確率 X については例 1 と同様であるが、これに加えて、偽を真と誤判定

6 機械学習利用システムと機械学習要素においてそれぞれの出力が異なる場合、利用時品質と外部品質も異なる。例えば、機械学習利用システムの出力から訓練データに関する情報を推定する攻撃を防ぐために、機械学習要素の出力の一部（判定結果の確からしさを示す数値〈確信度〉）を加工し、それを機械学習利用システムの出力とするケースがある。この場合、出力が異なることから、品質も異なる。

するタイプ2エラー（偽陽性）の確率Yについても、国籍の違いによる差が一定値以内に収まるべきという考え方に基づく要件例である。

- 例3（fairness through unawareness）：国籍を排除して訓練データ等を準備し、それを用いて機械学習要素を生成・テストすること。
  —— これは、訓練データ等の選択方法に着目した要件であり、配慮が必要な属性を、機械学習要素の生成プロセスから排除すればよいという考え方に基づく要件例である。ただし、国籍と強く相関する変数が採用されると、国籍の代理変数を使ったことになるリスクを排除できない（レッド・ライニング効果）。

- 例4（fairness through awareness あるいは individual fairness）：国籍を除く属性が類似している個人のペアにおいて、一定確率以上で判定結果が同一となること。
  —— これは、属性が類似している個人のペアであれば、融資判定結果も類似するはずであるという考え方に基づく要件例である。

このように、どの公平性の概念を採用するかによって利用時品質要件も異なる。ステークホルダーの意見を考慮しつつ、どれが適当かを検討する必要がある。また、例1、2、4のように定量的な指標を含む利用時品質要件の場合、指標の目標値を設定する必要もある。

ロ．属性間の関係の明確化

本節（1）イ．の例3の利用時品質要件を採用する場合、国籍と因果関係のある属性群を洗い出し、訓練データから除去する必要がある。

まず、収集対象の訓練データに含まれる属性をリストアップする。次に、国籍と他の属性との間の関係を分析し、一定の相関を有する属性を特定したうえで、それを訓練データから除去するか否かを検討する。国籍と一定の相関を有する属性であったとしても、それが融資判定上重要な要素であり使用しても差別につながらないと判断できる場合、その属性を除去しないという対応（no unresolved discrimination）もありうる。これは手続き的公正性を担保するというアプローチに相当する。

こうした検討を行ううえで、属性間の関係を視覚的に理解しやすいグラフ等で表現することが有効である。品質ガイドラインは、属性間の関係を比較的シンプルに表現することができるCausal Bayesian Network（CBN）を紹介している[7]。Chiappa

7 CBNは、ノード間の関係をベイジアン・ネットワークによって表現するグラフ。ノードが何らかの値を表し、ノード間を結ぶ辺が両端のノードの統計的な関係を表す。辺には向きが設定される非循環有向グラフである。

and Isaac [2019] は、CBN のグラフの各ノードにそれぞれ属性を割り当てるとともに、一定の因果関係が存在する属性（ノード）の間を辺で結ぶグラフを提案している。例えば、訓練データに国籍（属性 A）と勤務年数（属性 B）が含まれており、国籍が勤務年数に影響を及ぼしうる（potential cause）ことが判明した場合、国籍と勤務年数をノード A、B にそれぞれ割り当てたうえで、ノード A からノード B に対して矢印を描く。このような作業をリストアップしたすべての属性のペアに関して実施する。

なお、属性間の因果性が判明している場合、手続き的公正性が証明可能なケースとして以下のようなロジックが考えられる。ある属性 Z が融資判定結果と国籍の両方に因果性を持つ場合、この属性 Z を用いたとしても国籍の代理変数として利用されていることにはならない。結果として国籍が融資判定に影響力を持っているようなモデルになったとしても、見せかけの相関（関係性）であると主張することは可能である。ただし、因果性が適切に分析され、その説明が十分になされることによって手続き的公正性に関する合意をステークホルダーから得る必要があろう。

ハ．出力や訓練データに関する公平性メトリクスの選択と目標の設定

出力や訓練データに関する公平性メトリクスも利用時品質要件に基づいて設定する。

本節（1）イ．の例 1、2、4 の場合、利用時品質要件が定量的な目標値を含むことから、それが公平性メトリクスとなる。例 1 においては、融資可能と判定される確率の差分（絶対値）に数値目標を導入することで、融資判定に関する公平性メトリクスとすることができる。訓練データに関しては、利用時品質要件が国籍の融資判定への影響を小さくすることを要求していることから、定量的な指標ではないものの、例 3 と同じく、国籍も訓練データから排除することを目標として設定することが考えられる。こうした取扱いは例 1、2、4 のような公平性メトリクスと排他的ではなく、同時に適用することも可能である。

（2） システム開発フェーズ

システム開発フェーズでは、訓練データ等の収集、機械学習要素の生成・テストを実施する。

イ．訓練データ等の収集

訓練データ等の収集と準備に関して、品質ガイドラインは以下の要求事項を示している。

- 訓練データやテスト・データを、被覆性（網羅性）、均一性、妥当性に留意して収集する。
- 訓練データ等に関する公平性メトリクスを測定し、目標値が達成されているか否かを記録・確認する。
- 上記確認の結果、目標値との乖離が存在した場合には、それを解消するために、データに偏りが生じないようなデータ収集のプロセスや、データに内在する歪みを修正する手法[8] を適用することを検討する。
  —— データが所与であり変更が不可能なケースでは、データの準備以外の部分において、その乖離が出力に与える影響を最小化する方法を検討する。

ここでは、データの被覆性・均一性・妥当性の要求事項に焦点を当てて説明する。3 つの特性は以下のとおりである。

- 被覆性：想定されるシナリオやケースにそれぞれ対応するように、データの範囲や量が適切に準備されているか。
  —— 融資判定の場合、想定される借入申込のシナリオやケースを検討する。そのうえで、訓練データやテスト・データを収集し、想定したシナリオ等がカバーされているかを確認する。
- 均一性：訓練時のデータの分布が運用時の（想定される）入力の分布と整合しているか（偏りがないか）。
  —— 借入申込の各シナリオやケースの発生頻度を事前に検討し、発生頻度と整合的な量のデータを準備していることを確認する。
- 妥当性：データに誤りや不適切なものが含まれていないか。
  —— 訓練データのなかに、誤ったラベルが付けられたデータがないか、別のデータが混在していないかを確認する。

8 データの歪みを修正する主な手法として、品質ガイドラインは、Feldman *et al.* [2015] によって提案された disparate impact remover や、Calmon *et al.* [2017] によって提案された optimized pre-processing を紹介している。disparate impact remover は、配慮が必要な属性 X と相関を有し、機械学習要素の出力に影響を及ぼす属性 Y に関して、X が（Y を介して）出力に及ぼす影響（disparate impact）を小さくするように Y の分布を操作する手法である。optimized pre-processing は、元の訓練データと調整後の訓練データとの間の分布の差異を一定の量以下とすること（distortion control）等を制約条件とした

上記のうち、均一性が満たされず、あるシナリオやケースのサンプル数が運用時の想定分布に比べて少なかった場合、そのシナリオやケースの判定精度が低下しうる（sample size disparity）。また、訓練データにおける（配慮が必要な属性別の）データの構成に有意な差異がある場合、これが判定精度にギャップをもたらす可能性がある。品質ガイドラインは、訓練データの件数のバランスを属性別に調整する主な手法として reweighing を紹介している（Calders, Kamiran, and Pechenizkiy [2009]）。

【reweighing によるデータ件数の調整手順】

① 訓練データを、配慮が必要な属性の値（国籍 A、B）とラベル（融資可能／不可能）によってグループ化する（数値例は表 1 を参照）。
—— 例えば、グループ 1（国籍 A、融資可能）のデータ件数が 40、グループ 2（国籍 A、融資不可能）のデータ件数が 10、グループ 3（国籍 B、融資可能）のデータ件数が 20、グループ 4（国籍 B、融資不可能）のデータ件数が 30 とする。訓練データ全体のデータ件数は 100 とする。

② 各グループ（属性値とラベルのペア）が訓練データ全体に占める割合をそれぞれ求める。

③ 属性とラベルが互いに独立した事象であると仮定した場合の、各グループの発現確率（＝属性値の確率 × ラベルの確率）をそれぞれ求める。

④ 各グループにおいてウエイトを算出する。ウエイトは、属性とラベルが独立であると仮定した場合の各グループの発現確率（③で算出）を、訓練データ全体に占める各グループの割合（②で算出）で割った値。
—— ウエイトが 1 を超える場合、そのグループのデータが相対的に少ないと判断する。ウエイトが 1 未満である場合は、そのグループのデータが相対的に多いと判断。

⑤ 各グループのデータ件数を、ウエイトに応じて調整する。
—— 例えば、ウエイトが 2 となるグループ 2（国籍 A、融資不可能）はデータ件数を 2 倍に増やす。ウエイトが 3 分の 2 となるグループ 4（国籍 B、融資不可能）はデータ件数を 3 分の 2 に減らす。

この手法は、訓練データがサンプルの一部を利用したものであり、訓練データを増やすことが可能であること、また、訓練データを減らしても問題にならない程度十分に訓練データが存在することが前提になっている。また、本事例では、訓練に

うえで、disparate impact を最小化するように訓練データを調整する手法である。

表 1　Reweighing の数値例

グループ			データ件数	データセットにおける発現確率（PX）	属性とラベルが独立と仮定したときの発現確率（PY）	ウエイト（PY/PX）	ウエイトに基づくデータ件数の調整の方法
番号	国籍（属性）	融資判定結果（ラベル）					
1	A	可能	40	2/5 (=40/100)	3/10 (=50/100 × 60/100)	3/4 (=3/10 ÷ 2/5)	件数を 3/4 に減らす
2	A	不可能	10	1/10 (=10/100)	1/5 (=50/100 × 40/100)	2 (=1/5 ÷ 1/10)	件数を 2 倍に増やす
3	B	可能	20	1/5 (=20/100)	3/10 (=50/100 × 60/100)	3/2 (=3/10 ÷ 1/5)	件数を 1.5 倍に増やす
4	B	不可能	30	3/10 (=30/100)	1/5 (=50/100 × 40/100)	2/3 (=1/5 ÷ 3/10)	件数を 2/3 に減らす

資料：Calders, Kamiran, and Pechenizkiy [2009]

使われたデータにおいて融資判定が適切であったかについては考察していない。下された判定をそのまま学習することを想定している。

ロ．機械学習要素の生成・テスト

機械学習要素を生成し、公平性に関するテストを実施する。訓練済みの機械学習要素にテスト・データを入力し、それらに対する出力を得て公平性メトリクスを算出する。算出された公平性メトリクスと目標値を比較し、公平性メトリクスが目標値を達成したか否かを確認する。達成していない場合、訓練済み機械学習モデルに何らかの対応を行う必要がある。そうした場合の対応について、品質ガイドラインは以下を示している。

- 訓練済み機械学習モデルの生成中あるいは生成後に対処する手法を適用し、目標値との乖離の解消を図る。
- 上記により乖離を解消できない場合、運用時における対応を検討する。

（イ）　訓練済み機械学習モデルの生成中に対処する手法

品質ガイドラインでは、Kamishima *et al.* [2012] による prejudice remover regularizer、Beutel *et al.* [2017] や Zhang, Lemoine, and Mitchell [2018] による adversarial debiasing を主な手法として紹介している。

【prejudice remover regularizer】

・　本手法は、配慮が必要な属性（例えば国籍）と出力（融資判定結果）との間

の相関係数（正確には相互情報量）[9] を正規化項（regularizer）として損失関数に加えて訓練するものである。

・ 訓練では、損失関数の値を最小化するパラメータを探索する。その際、国籍と融資判定結果の相関が大きくなるように探索が進むと、正規化項が大きくなり損失関数の値も上昇することから、相関が強くなる方向の探索が抑制される。

【adversarial debiasing】

・ 融資判定結果から配慮が必要な属性の値を推定することが困難であれば、その判定結果は配慮が必要な属性の影響を受けていないとみなしてよいかもしれない。本手法はこうしたアイデアに基づく手法であり、Generative Adversarial Network（GAN）[10] を活用している。

・ 具体的には、融資可否を判定する機械学習要素（判定器）と、その判定器の出力から（対応する入力に含まれる）配慮が必要な属性（国籍）を予測する機械学習要素（予測器）を訓練する。最終的に、融資判定結果から配慮が必要な属性の値を予測することが困難であるという状態を達成する判定器を生成する。

（ロ） 機械学習要素の出力に調整を施す手法

訓練済み機械学習モデルを生成する過程での対応が困難なケースでは、機械学習要素の出力に調整を施す手法を適用する。品質ガイドラインは、主な手法として、Hardt, Price, and Srebro [2016] の手法や Kamiran, Karim, and Zhang [2012] の手法（reject option of probabilistic classifier: ROC）を示している。ただし、これらはいずれも融資判定結果の精度低下につながる可能性がある。

【Hardt, Price, and Srebro [2016] の手法】

・ 配慮が必要な属性ごとに異なる判定閾値を割り当てる。判定閾値は、その値

9 属性と出力の相関係数として相互情報量（mutual information）を用いる。相互情報量は、モデルへの入力（属性）の値を知ったときに、その出力の値について得られる情報の量と定義される。確率論や情報理論で用いられる統計量である。

10 GAN は、2 つの機械学習要素（例えば D と G）を互いに競わせることにより、双方の精度を同時に高め、最終的に所望の D、G を得る手法である（Goodfellow *et al.* [2014]）。D は何らかの入力を与えられてそれに関する判定を行うモデル（discriminative model）であり、G は D の判定の対象となる入力を生成するモデル（generative model）である。まず、G に対して、所与のデータ X を入力として与え、D の判定をより難しくするように一定のノイズを付加してデータ Y を出力させる。次に、D に対して、Y を入力として与え、Y が X（と同じ分布のサンプル）か否かを判定させる。判定結果を用いて D と G をそれぞれ再訓練する。この一連の処理を繰り返すことによって、D と G の精度を同時に向上させる。本文の事例では、判定器が D に、予測器が G にそれぞれ相当する。

よりも高い確信度[11]が得られた場合に融資可能と判定する値を指す。

・ 例えば、国籍 A の個人への融資判定の確信度が国籍 B の個人の融資判定の確信度よりも高いとする。この場合、国籍 B の個人の判定閾値を国籍 A のそれよりも低く設定する。これにより、国籍 B の個人が融資可能と判定される確率が高まり、公平性メトリクスを改善する効果が期待される[12]。

【ROC】

・ 判定結果の確信度が低い場合、その判定をあいまいと評価し使用すべきでないと判断する手法である。ただし、判定精度の低下とのトレードオフ関係を考慮しつつ確信度の足切りラインをどのように設定するかといった課題がある（Caton and Haas [2020]）。

・ 融資不可能との判定結果の確信度が足切りラインを下回りその判定を使用しないケース（reject option）では、その入力に対応する国籍（配慮が必要な属性）を確認する。確認の結果、その国籍のグループにおける判定結果が融資不可能に偏っていた場合には、融資不可能との判定結果を融資可能に変更することを検討する。こうした変更をどの程度行うかはその融資案件の内容を人間が精査して個別に決定することになる。

(3) システム運用フェーズ

このフェーズでは、サービス提供者は、システムの稼働状況（融資判定結果を含む）をモニターするとともに、不適切な判定結果の出力といった問題が発生した際に、その判定結果の修正や機械学習要素の更新を実施する。その前提として、不適切な状態を定義してそれを検知する手法を確立しておく必要がある。品質ガイドラインやガバナンス・ガイドラインは、それぞれ機械学習利用システムへの要求事項や不適切性に関する確認項目例を示している。

品質ガイドラインは、運用時に得られたデータを訓練データとして活用して機械学習要素を随時更新するケース（active learning）において、判定結果が不適切であった場合、訓練によって品質が逆に低下しうることを説明している。そのうえで、品質低下を防止するために判定結果に関して公平性メトリクスを測定すること

11 確信度（confidence value）は、予測・推論結果の確からしさを示す値（0 と 1 の間の値）である。ここでは、予測・推論結果が融資可能の確信度として表現され、その値が判定閾値よりも大きい場合、融資可能という判定結果を出力するというケースを想定している。例えば、確信度が 0.7 であった場合、判定閾値が 0.7 よりも小さな値であったならば、融資可能との判定結果を出力する。

12 ただし、判定時の誤りを最小化するようにして得られた判定閾値を変更することになるため、融資判定の精度が低下する可能性がある。これへの対応として、Pleiss *et al.* [2017] は判定閾値を確率的に決定する手法を提案しており、精度の低下を緩和できるとしている。

を要求している。これは、新たな訓練データのなかに目標値に合致しないものが入り込むのを回避し、機械学習要素の更新に不適切なデータを使用しないようにするためである。

また、判定結果が公平性の観点から許容される範囲を超えるケースがありうる。ガバナンス・ガイドラインは、こうしたケースを認識する重要性を指摘し、以下の趣旨の確認項目例を示している（補論 2 を参照）。

- 公平性の指標（公平性メトリクス）を用いる場合、許容範囲を超えた出力に対して警告を発する措置を講じたか。
- 許容範囲を超える出力の可能性やその際の対応をサービス利用者に予め留意事項として伝え理解を得るようにしたか。
- システムの出力に問題が生じた際に、AI を用いないプロセスに変更するといった仕組みを整えているか。
- 人間の主体的な関与の機会を適切に提供しているか。

このように、機械学習要素が目標値を達成しない判定結果を出力する可能性があるのであれば、そうした特性をサービス利用者に予め伝えて理解を得るという方法を示している。また、判定結果に問題が生じた際の備えとして、機械学習要素の出力結果を用いずに人間（サービス提供者）が判定をやり直すといった対応を準備しておくことを挙げている。

5. 公平性の要件の設定にかかる課題

3、4 節で説明したように、公平性に配慮した機械学習利用システムを実現する際に原則・社会規範、各種ガイドラインを活用し、公平性に関する要求事項が達成されているかを検証できるようになっている。もっとも、技術のみでは必ずしも解決できない部分が残されている。本節では、そうした課題として、①公平性に関するサービス要件を設定する際に、ステークホルダーの期待や意見をどのようにして把握するか、②さまざまな公平性の概念のなかでどれを選択するかに焦点を当て、最近の研究成果を紹介しつつ、今後の課題を述べる。

(1) ステークホルダーの期待や意見をどのように把握するか

公平性をどう捉えるかは、個人によって、また、ステークホルダーの立場によっても異なりうる。そのため、サービス利用者となりうる個人や企業の期待や意見をサービス要件に適切に反映できるよう対応することが求められる。

イ. ステークホルダーとの対話

Stumpf *et al.* [2021] は、ステークホルダーと実際に対話しながら公平性に関する意見や判断の基準を抽出し、それを機械学習利用システムの設計に反映させるアプローチ（co-design）に基づく手法を提案している。Stumpf *et al.* [2021] は、個人ローン審査に用いる機械学習利用システムを設計する場面を想定し、上記手法によってステークホルダーの意見や判断の基準を抽出するためのケース・スタディを実施した。このケース・スタディではワークショップが開催され、ステークホルダーとして、ローンの申込者となりうる一般の個人、金融機関の融資担当者、データ・サイエンティスト（機械学習利用システムの構築を担当）が参加した[13]。各参加者は、主催者が用意した質問等を手掛りに対話を行い、公平性の定義や過去に公平（または不公平）と感じた経験等を述べた。また、ローン申込とそれに対する融資の可否について複数の例が示され、それぞれの例に関して、決定は『公平』か、決定の公平性を評価する際にどのような情報を用いたかといった問いに回答した。こうして得られた情報から公平性の判断基準が導出された。

このような手法はステークホルダーへの説明や納得感の醸成という観点でも重要である。サービス提供者が、サービス要件等を決定する際にステークホルダーの一部や代表者と適切なコミュニケーションを行った旨を、具体的な公平性のサービス要件および利用時品質要件とともにサービス利用者に説明することができれば、サービス利用者の納得感や機械学習利用システムの動作に対する信頼が高まると期待できる。

ロ. 課題

もっとも、ワークショップ等を開催するためには、サービス利用者やシステム開発者の参加者を募ったうえで、相応の時間をかけて、対象としているサービスの内容を各参加者に十分理解してもらう必要がある。また、日常で公平あるいは不公平と感じた経験を参加者から聴取し、これを公平性に関するシステム要件にどう織り込むかを検討することも必要である。その他の検討すべき事項として、参加者をど

13 ワークショップは、一般の個人（12 名参加）を対象とするものと、融資担当者（6 名）およびデータ・サイエンティスト（6 名）を対象とするものがそれぞれ 2 回ずつ行われた。これらのワークショップの詳細については、それぞれ Nakao *et al.* [2022a, b] を参照されたい。

のように選択しリストアップするか、参加者の人数をどうするか、ユーザ・インタフェースをどのように設定するかといった項目が挙げられる。

公平性に関するステークホルダーの期待や意見をサービス要件に反映することは重要な検討課題である。しかし、機械学習における公平性に関するノウハウや知見が主催者側に足りない場合には、ステークホルダーの期待や意見をどのようにサービス要件に反映すればよいかを判断することは難しい。

こうした問題に対しては、類似のサービスの提供を検討する複数の金融機関や組織が共同でワークショップを開催し、得られた公平性に関する期待や意見、判断基準を共有・利用するという方法が考えられる。また、ノウハウや知見の不足に対しては、公平性の問題に詳しいデータ・サイエンティストの協力を仰ぎつつ共同で研究を行うといった対応によって、金融機関がノウハウの蓄積を中長期的に進めることが有用であろう。

本節（1）イ．において紹介したような手法に関する研究開発が今後さらに進むことが期待される。特に、ステークホルダーとのコミュニケーションを一段と円滑にするインタフェースやステークホルダーの発言を分析するツールの研究開発が注目される。

(2)　どの公平性の概念を選択するか

利用時品質要件を特定する際にどのような公平性の概念を採用するかが問題となる。実際に、各種の公平性の概念のうち、どれが満たされるものなのか。ここでは、融資判定向けの機械学習要素に関して公平性メトリクスを測定した Verma and Rubin [2018] の研究を紹介する。

イ．16 種類の公平性メトリクスの測定

Verma and Rubin [2018] は、German Credit Dataset の約 1,000 件の個人向けローン申込のデータ[14] を使用し、融資可否を判定する機械学習要素（ロジスティック回帰モデル）を生成したうえで、テスト・データにおける判定結果に基づいて 16 種類

14 German Credit Dataset は、ドイツ・ハンブルク大学のハンス・ホフマン教授によって構築された個人向けローン申込のデータベースであり、UCI Machine Learning Repository によって提供されている。次の 20 種類の属性データが含まれている。①借入金額、②借入期間、③借入金の用途、④小切手口座の状態、⑤預貯金口座の状態、⑥既存の借入金額、⑦これまでの借入金の履歴、⑧分割返済計画、⑨分割返済対象の金額、⑩所有財産、⑪持ち家の有無、⑫持ち家の年数、⑬電話（の有無）、⑭勤務状態、⑮勤続年数、⑯本人の状態と性別（属性値は、独身男性、既婚男性、既婚あるいは離婚女性、離婚男性の 4 つ）、⑰年齢、⑱外国人労働者か否か、⑲扶養家族の人数、⑳他の債務を負っているか否か（other debtors）。各レコードには、ラベルとして融資可能（good）あるいは不可能（bad）が含まれている。

の公平性メトリクスをそれぞれ測定した。

ここで、公平性メトリクスの測定は、訓練データにおける 20 種類の属性のうち配慮が必要な属性として性別に焦点を当てて、性別が融資判定結果にどのような影響を与えるかという観点で行われた。データセットでは、性別にかかわる属性として、独身男性、既婚男性、既婚あるいは離婚女性、離婚男性の 4 つが存在していた。本研究では、性別の違いが判定結果に与える影響を分析することを目的としていたため、既婚男性と離婚男性のグループを合わせて、既婚あるいは離婚男性のグループを作成し、既婚あるいは離婚女性のグループと比較した（独身男性のデータを使用していない）。したがって、この研究における男女の比較は、既婚あるいは離婚の男女の比較となっている。

表 2 に示す公平性の概念を設定し、上記データセットに対して公平性メトリクスを測定した。その結果、半数の公平性の概念において、グループ間で公平性メトリクスに有意な差がみられた。

例えば、公平性の概念として demographic parity を採用した場合、融資可能と判定される確率を各グループで算出し（女性のグループは 0.75、男性のグループは 0.81）、両者の差分を計算した（差分は 0.06）。その結果、この確率の差分が有意であることが判明し、公平性の目標を達成できなかったと評価した。一方、conditional statistical parity を採用した場合には、融資可能と判定される確率の（男女グループ間での）差分が 0.03 となった。この差分については有意であるとはいえず、公平性の目標を達成することができたと評価した。なお、fairness through unawareness を採用した場合には、機械学習要素の生成時に性別の属性を使用しなかったことから、公平性の目標を達成することができたと評価した。

表 2 の検証とは別に、本研究で使用した機械学習要素において、性別が出力に対して影響を与えていないことをテスト・データに関して確認した。各テスト・データ（性別の属性値を含む）における性別の属性値を変更し（女性を男性に、男性を女性に変更）、変更後のテスト・データを機械学習要素に入力した。その結果、性別の値を変更したテスト・データに対する出力は変更前のテスト・データに対する出力と同一となり、性別の値が変化したとしてもテスト・データに対する出力はその影響を受けなかったことが示された。

ロ. 課題

上記の研究事例では、配慮が必要な属性として性別に焦点を当てていたが、実際のシステム開発・運用の場面では、性別だけでなく、複数の属性を配慮が必要な属性と捉え、利用時品質要件もそれぞれの属性に対して設定するケースが想定される。このように、複数の属性を同時に取り扱うケースにおいて公平性の概念をどのように満たすことができるかについては明確になっておらず、今後の課題といえる。

表２　【研究事例】ローン申込における性別に関する公平性メトリクス

公平性の概念	意味〈評価：「○」は達成、「×」は未達〉	公平性メトリクスの測定値
demographic parity (group fairness)	融資可能と判定される確率が女性と男性のグループ間で同じになる（×）	女性：0.75 男性：0.81
conditional statistical parity	融資判定で重視すべき属性（借入金額、借入履歴、勤務状況、年齢）を同一としたサブグループを男女別に抽出したとき、融資可能と判定される確率が男女間で同じ（○）	女性：0.49 男性：0.46
predictive parity	融資可能判定の正解確率（実績と一致する確率）が男女間で等しい（○）	女性：0.74 男性：0.73
predictive equality	融資不可能実績を融資可能と誤判定した確率（偽陽性率）が男女間で等しい（×）	女性：0.55 男性：0.70
equal opportunity	融資可能実績を融資不可能と誤判定した確率（偽陰性率）が男女間で等しい（○）	男女とも 0.14
equalized odds	①融資可能実績を融資可能と判定した確率と、②融資不可能実績を融資可能と誤判定した確率（偽陽性率）が、それぞれ男女間で等しい（×）	①男女とも 0.86 ②女性：0.55 男性：0.70
conditional use accuracy equality	① predictive parity に加えて、②融資不可能実績を正しく融資不可能と判定した確率が男女間で等しい（×）	①女性：0.74 男性：0.73 ②女性：0.63 男性：0.49
overall accuracy equality	判定結果全体を対象とし、これらが実績と一致する確率が男女間で等しい（○）	女性：0.71 男性：0.68
treatment equality	偽陽性率に対する偽陰性率の比が、男女間で等しい（×）	女性：0.62 男性：0.56
test fairness	判定の確信度別に 11 グループに分け、各グループ別に融資可能と判定される確率が、男女間で等しい（○）	11 のグループのうち 6 つで確率が一致
well calibration	判定の確信度別に 11 グループに分け、各グループ別に融資可能と判定される確率が、男女間で等しく、かつ、確信度と一致する（○）	11 のグループのうち 4 つで確率が一致
balance for positive class	融資可能実績における確信度が男女間で等しい（○）	男女とも 0.72
balance for negative class	融資不可能実績における確信度が男女間で等しい（×）	女性：0.52 男性：0.61
causal discrimination	同一の属性（男女以外）を有する申込への判定は必ず同一である（×）	同一の属性をもつ男女のうち判定結果が一致する確率は 0.912
fairness through awareness	類似の属性を有する申込への判定は男女属性によらず類似する（×）	
fairness through unawareness	性別を除いた属性を訓練データとして機械学習要素を生成する（○）	――

資料：Verma and Rubin [2018]

また、実際のサービス運用においては、公平性メトリクスが目標値を達成する属性と達成できない属性が混在する状況が発生するかもしれない。こうした状況への対応は、機械学習利用システムのシステム開発のアセスメント・フェーズにおいて

検討することになる。技術的な対応に限界があるという場合には、目標の達成度合いが異なる属性が混在する状態でシステムを利用した際に想定されうるリスク・シナリオを検討し、必要に応じて開示することが求められる。実際に公平性メトリクスの目標値が達成されないケースに対応する際には、人間による判定への切替えが考えられることから、そのための準備が必要となる。サービスに対する顧客からの信頼を維持していくうえで重要な検討課題といえよう。

参考文献

荒井ひろみ・Urlich Aïvodj・Olivier Fortineau・Sébastien Gambs・原　聡・Alain Tapp、「機械学習の説明における公正さの偽装」、第 34 回人工知能学会全国大会論文集、人工知能学会、2020 年（https://www.jstage.jst.go.jp/article/pjsai/JSAI2020/0/JSAI2020_3N5OS11b04/_pdf/-char/ja、2023 年 9 月 8 日）

井上紫織・宇根正志、「金融分野で活用される機械学習システムのセキュリティ分析」、『金融研究』第 39 巻第 1 号、日本銀行金融研究所、2020 年、17〜48 頁

宇根正志・清藤武暢、「機械学習システムにおけるソフトウェアの品質評価の現状と課題」、『金融研究』第 39 巻第 1 号、日本銀行金融研究所、2020 年、49〜74 頁

金融庁、「金融サービス業におけるプリンシプルについて」、金融庁、2008 年（https://www.fsa.go.jp/news/19/20080418-2/01.pdf、2023 年 9 月 7 日）

国立研究開発法人産業技術総合研究所、「機械学習品質マネジメントガイドライン 第 3 版（Revision 3.2.1）」、国立研究開発法人産業技術総合研究所、2022 年（https://www.digiarc.aist.go.jp/publication/aiqm/AIQuality-requirements-rev3.2.1.0079-signed.pdf、2023 年 9 月 7 日）

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」、内閣府、2019 年（https://www8.cao.go.jp/cstp/aigensoku.pdf、2023 年 9 月 8 日）

丸山　宏、「機械学習工学に向けて」、日本ソフトウェア科学会第 34 回大会（2017 年度）講演論文集、日本ソフトウェア科学会、2017 年（http://jssst.or.jp/files/user/taikai/2017/GENERAL/general6-1.pdf、2023 年 9 月 8 日）

AI 原則の実践の在り方に関する検討会、「AI 原則実践のためのガバナンス・ガイドライン Ver. 1.1」、経済産業省、2022 年（https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_1.pdf、2023 年 9 月 8 日）

AI プロダクト品質保証コンソーシアム、「AI プロダクト品質保証ガイドライン 2023.06 版」、AI プロダクト品質保証コンソーシアム、2023 年（https://www.qa4ai.jp/QA4AI.Guideline.202306.pdf、2023 年 9 月 7 日）

Agarwal, Sray, and Shashin Mishra, *Responsible AI: Implementing Ethical and Unbiased Algorithms*, Springer, 2021.

Beutel, Alex, Jilin Chen, Zhe Zhao, and Ed H. Chi, “Data Decisions and Theoretical Implications When Adversarially Learning Fair Representations,” arXiv: 1707.00075v2, 2017.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings,” Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS), Association for Computing Machinery, 2016, pp. 4356–4364 (available at https://dl.acm.org/doi/pdf/10.5555/3157382.3157584、2023 年 9 月 8 日).

Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy, “Building Classifiers with Independency Constraints,” Proceedings of 2009 IEEE International Conference on Data Mining Workshops, IEEE, 2009, pp. 13–18 (available at https://www.win.tue.nl/~mpechen/publications/pubs/CaldersICDM09.pdf、2023 年 9 月 8 日).

———, and Sicco Verwer, “Three Naive Bayes Approaches for Discrimination-Free Classification,” *Data Mining and Knowledge Discovery*, 21, Springer, 2010, pp. 277–292 (available at https://link.springer.com/content/pdf/10.1007/s10618-010-0190-x.pdf、2023 年 9 月 8 日).

Calmon, Flavio P., Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney, “Optimized Pre-Processing for Discrimination Prevention,” Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Association for Computing Machinery, 2017, pp. 3995–4004 (available at https://dl.acm.org/doi/pdf/10.5555/3294996.3295155、2023 年 9 月 8 日).

Caton, Simon, and Christian Haas, “Fairness in Machine Learning: A Survey,” arXiv: 2010.04053v1, 2020.

Chiappa, Silvia, and William S. Isaac, “A Causal Bayesian Networks Viewpoint on Fairness,” in E. Kosta, J. Pierson, D. Slamanig, S. Fischer-Hübner, and S. Krenn, eds. *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data,* Springer Nature, 2019, pp. 3–20.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, “Fairness through Awareness,” Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Association for Computing Machinery, 2012, pp. 214–226 (available at https://www.cs.toronto.edu/~toni/Papers/awareness.pdf、2023 年 9 月 8 日).

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, “Certifying and Removing Disparate Impact,” Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, 2015, pp. 259–268.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, “Predictably Unequal? The Effects of Machine Learning on Credit Markets,” *The Journal of Finance*, 77(1), 2022, pp. 5–47.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative Adversarial Nets,” Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), 2, Association for Computing Machinery, 2014, pp. 2672–2680 (available at https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf、2023 年 9 月 8 日).

Hardt, Moritz, Eric Price, and Nathan Srebro, "Equality of Opportunity in Supervised Learning," Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS), Association for Computing Machinery, 2016, pp. 3323–3331 (available at https://dl.acm.org/doi/pdf/10.5555/3157382.3157469、2023 年 9 月 8 日).

Hurley, Mikella, and Julius Adebayo, "Credit Scoring in the Era of Big Data," *Yale Journal of Law & Technology*, 18, 2016, pp. 148–216.

Kamiran, Faisal, Asim Karim, and Xiangliang Zhang, "Decision Theory for Discrimination-Aware Classification," Proceedings of the 2012 IEEE 12th International Conference on Data Mining, IEEE, 2012, pp. 924–929 (available at https://web.lums.edu.pk/~akarim/pub/decision_theory_icdm2012.pdf、2023 年 9 月 8 日).

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma, "Fairness-Aware Classifier with Prejudice Remover Regularizer," *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, *Lecture Notes in Computer Science*, 7524, Springer, 2012, pp. 35–50.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, 54 (6), Article No. 115, 2021, pp. 1–35.

Nakao, Yuri, Lorenzo Strappelli, Simone Stumpf, Aisha Naseer, Daniele Regoli, and Giulia Del Gamba, "Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness," *International Journal of Human-Computer Interaction*, 39(9), 2022a, pp. 1762–1788.

———, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli, "Towards Involving End-Users in Interactive Human-in-the-Loop AI Fairness," *ACM Transactions on Interactive Intelligent Systems*, 12(3), Article No. 18, 2022b, pp. 1–30.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger, "On Fairness and Calibration," Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Association for Computing Machinery, 2017, pp. 5684–5693 (available at https://dl.acm.org/doi/pdf/10.5555/3295222.3295319、2023 年 9 月 8 日).

Rea, Stephen C., "A Survey of Fair and Responsible Machine Learning and Artificial Intelligence: Implications of Consumer Financial Services," Social Science Research Network, 2020 (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3527034、2023 年 9 月 8 日).

Stumpf, Simone, Lorenzo Strappelli, Subeida Ahmed, Yuri Nakao, Aisha Naseer, Giulia Del Gamba, and Daniele Regoli, "Design Methods for Artificial Intelligence Fairness and Transparency," Joint Proceedings of the ACM IUI Workshops, Association for Com-

puting Machinery, 2021 (available at https://eprints.gla.ac.uk/261474/1/261474.pdf、2023 年 9 月 8 日).

Verma, Sahil, and Julia Rubin, “Fairness Definitions Explained,” Proceedings of the International Workshop on Software Fairness, Association for Computing Machinery, 2018, pp. 1–7 (available at https://fairware.cs.umass.edu/papers/Verma.pdf、2023 年 9 月 8 日).

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell, “Mitigating Unwanted Biases with Adversarial Learning,” Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, 2018, pp. 335–340 (available at https://dl.acm.org/doi/pdf/10.1145/3278721.3278779、2023 年 9 月 8 日).

補論 1. AI 社会原則における公平性の原則

AI 社会原則には、AI の研究開発や社会実装における基本理念、考慮すべき問題、基本原則が示されている。「AI の適切で積極的な社会実装を推進するためには、各ステークホルダーが留意すべき基本原則を定めることが重要である」として、①人間中心、②教育・リテラシー、③プライバシー確保、④セキュリティ確保、⑤公正競争確保、⑥公平性、説明責任および透明性、⑦イノベーションに関する原則が定められている。このうち、公平性の原則は、AI 社会原則の 4.1 節（6）「公平性、説明責任及び透明性の原則」に以下のとおり記載されている。

【AI 社会原則 4.1 節（6）「公平性、説明責任及び透明性の原則」の引用】
「AI-Ready な社会」においては、AI の利用によって、人々が、その人の持つ背景によって不当な差別を受けたり、人間の尊厳に照らして不当な扱いを受けたりすることがないように、公平性及び透明性のある意思決定とその結果に対する説明責任（アカウンタビリティ）が適切に確保されると共に、技術に対する信頼性（Trust）が担保される必要がある。

- AI の設計思想の下において、人々がその人種、性別、国籍、年齢、政治的信念、宗教等の多様なバックグラウンドを理由に不当な差別をされることなく、全ての人々が公平に扱われなければならない。
- AI を利用しているという事実、AI に利用されるデータの取得方法や使用方法、AI の動作結果の適切性を担保する仕組みなど、用途や状況に応じた適切な説明が得られなければならない。
- 人々が AI の提案を理解して判断するために、AI の利用・採用・運用について、必要に応じて開かれた対話の場が適切に持たれなければならない。
- 上記の観点を担保し、AI を安心して社会で利活用するため、AI とそれを支えるデータないしアルゴリズムの信頼性（Trust）を確保する仕組みが構築されなければならない。

上記の 4 項目のうち、1〜3 番目の項目は、それぞれ公平性、説明責任、透明性の原則であり、4 番目は各特性に共通した原則と捉えることができる。

補論 2. ガバナンス・ガイドラインにおける公平性に関する評価項目

ガバナンス・ガイドラインでは、機械学習利用システムの開発・運用の管理を、一定の手続きに沿って実施することが紹介されている。こうした管理は AI マネジメント・システムと呼称されている。また、システム運用者（サービス提供者に対応）あるいはシステム開発者がシステムの機能や特性の達成度合いを評価する[15]ための項目例も示されている。それらのうち、公平性に関する評価項目例と確認項目例をまとめると表 A-1 の 7 項目となる。評価項目例に沿って評価を行った結果、公平性に関するリスクが許容範囲を超えているならば対応を行うといった活用法が考えられる。

15 より具体的には、設計段階で設定された機能・特性に関する目標が、実際のシステムにおいて達成されているか否かを評価するものである。

表 A-1　公平性に関する主な評価項目例と確認項目例

評価項目例	具体的な確認項目の例（要約）
①システムに求められる公平性を把握しているか	・公平性に関するインシデント事例を調査したか ・サービス対象地域、また、類似のシステムにおいて、偏見や差別的な扱いが生じたといった指摘（の有無）を確認したか ・公平性に関する適切な指標を選択し、その指標に基づいて公平性に関する許容範囲を評価したか
②システムの公平性に関する課題に対処したか	・可能な範囲で、開発チームのスタッフの多様性を高めたか ・可能な範囲で、サービス対象地域の規制、慣習、商慣行等を理解するスタッフを開発チーム等に加えたか ・公平性の指標を用いる場合、許容範囲を超えた出力に対して警告を発する措置を講じたか ・許容範囲を超える出力の可能性やその際の対応をサービス利用者に予め留意事項として伝え理解を得るようにしたか
③人間によるシステムへの主体的な関与の機会を確保したか	・公平性向上の観点から、人間による制御可能性を含め、人間の主体的な関与の機会の必要性を検討したか ・システムの出力の採否や中断・停止を決定する自由や機会をサービス提供者に提供する設計としているか ・システムの出力に問題が生じた際に、AI を用いないプロセスに変更するといった仕組みを整えているか ・サービス提供時にシステムに過度に依存しない設計方針か
④データセット設計時に差別を維持・助長しないよう配慮したか	・特定の属性に基づく差別が存在し、これがデータセットに反映されていた場合、システムが差別を取り込んでしまう（現実を再現してしまう）危険性について検討したか ・差別を維持・助長するデータセットの不使用を再現性より優先したか ・上記事態を回避するために、データセット設計スタッフの多様性を高めたか
⑤システムの公平性を確保したか	・特定の属性に基づく差別を維持・助長していないかを評価したか ・各属性の出力への影響度や感度を評価したか ・公平性の定義や指標を調査し、適切な定義や用途が存在する場合には、それらを用いて公平性を客観的に評価したか
⑥公平性に関する課題への対処について理解したか	・公平性の指標に基づく警告発生措置を講じた場合、指標やその警告の意味を適切に理解しているか ・公平性に関する（システムの開発主体による）注意事項を理解したか
⑦人間の主体的な関与の機会を提供しているか	・人間の主体的な関与の機会の重要性や、サービス提供者が意思決定に際してシステムに過剰に依存しない設計の重要性を、その理由も含めてサービス提供者が適切に理解したか ・サービス提供者は、サービス利用者に対して人間の主体的な関与の機会を適切に提供しているか

資料：AI 原則の実践の在り方に関する検討会［2022］

補論 3. 品質ガイドラインにおける公平性確保に向けた要求事項

品質ガイドラインは、機械学習利用システムの外部品質として、公平性、リスク回避性（危害・危険回避性、安全性）、AI パフォーマンス（有効性）、プライバシー、AI セキュリティを取り上げている。公平性を確保するうえでの要求事項は、法令・規則による要請の有無等に基づいて以下の 3 つのレベル（AIFL 0〜2）に分けられている。

- AIFL 2：法令・規則・社会的なガイドライン等によって公平性の実現が要請されているケース
- AIFL 1：法令・規則等による要請はないものの、機械学習利用システムの出力が公平であることを説明できないならば、サービスの社会受容性等が影響を受けたり、障害が発生したりするケース
- AIFL 0：上記以外のケース

金融分野では、金融庁による「金融サービス業におけるプリンシプルについて」においてサービス利用者の公平な取扱いが原則とされている。これを社会的なガイドラインと捉えると、金融サービスに用いられる機械学習利用システムは、「公平性の実現が要請されている」として AIFL 2 のケースに相当する。

補論 4. さまざまな公平性の概念

公平性の概念にはさまざまなものがある。Verma and Rubin [2018] は、これらを①統計値に基づくもの、②属性の類似性に基づくもの、③属性・出力間の関係性に基づくものに分類している。以下では、この分類に沿って説明する。

(1) 統計値に基づく概念

統計値に基づく概念（表 A-2 を参照）のうち、特に研究論文でよく取り上げられるのが demographic parity、equal opportunity、equalized odds である。demographic parity は、「配慮が必要な属性（例えば国籍）の値に応じて入力をグループ分けしたときに、出力がポジティブ（例えば融資可能）となる確率がグループ間で等しい」という状態を公平と捉える。これは、入力をグループ分けするとともに、予測・推論結果に着目した概念といえる。equal opportunity は、「配慮が必要な属性の値に応じて入力をグループ分けしたとき、ラベル（実際のデータ）がポジティブである入力に対する出力（判定結果）がネガティブ（例えば融資不可）となる確率が、グループ間で等しい」という状態を公平と捉える。これは、予測・推論結果と入力の両方に着目した概念といえる。equalized odds は、「配慮が必要な属性の値に応じて入力をグループ分けしたとき、①ラベルがポジティブであるときに、出力が実際にポジティブとなる確率、および、②ラベルがネガティブであるときに、出力がポジティブとなる確率が、それぞれグループ間で等しい」という状態を公平と捉える。

なお、equality（平等、結果が一様に等しい）は公平、公正とは異なる概念であるが、ここでは確率等の数値・数量が等しい（equal）という意味で公平性の概念の定義に使われている。

(2) 属性の類似度に基づく概念

この分類の概念は、「配慮が必要な属性以外の属性が類似している個人間では、予測・推論結果も類似したものとするべきである」という考え方に基づく。Dwork *et al.* [2012] は、配慮が必要な属性に着目してグループ分けし、グループ間の差異を解消する考え方に基づく概念（例えば、equal opportunity）の問題点を指摘している。この概念を用いると、グループ内の個人の差異をサービス内容に反映できない場合がある。より望ましい概念として、配慮が必要な属性の影響を排除しつつ、その他の属性もサービス内容に反映できるものを検討すべきとの見方を示している。

表 A-2　統計値に基づく主な公平性の概念

公平性の概念	意味
overall accuracy equality	予測・推論結果がラベル（ポジティブ／ネガティブ）と一致する確率が比較対象グループ間（以下、グループ間）で等しい
treatment equality	偽陽性率（ネガティブをポジティブと誤認する確率）に対する偽陰性率（ポジティブをネガティブと誤認する確率）の比がグループ間で等しい
equal opportunity	偽陰性率がグループ間で等しい ― false negative error rate balance とも呼ばれる
equalized odds	ラベルがポジティブである入力において、出力もポジティブとなる確率、および、偽陽性率が、それぞれグループ間で等しい ― equality of odds、conditional procedure accuracy equality とも呼ばれる
predictive equality	偽陽性率がグループ間で等しい ― false positive error rate balance とも呼ばれる
predictive parity	出力がポジティブである入力のラベルもポジティブである確率がグループ間で等しい
conditional use accuracy equality	predictive parity に加えて、出力がネガティブである入力のラベルがネガティブである確率もグループ間で等しい
demographic parity	出力がポジティブとなる確率がグループ間で等しい ― statistical parity、statistical fairness とも呼ばれる
conditional statistical parity	各グループから一部の特定の属性の値が同一となるサブグループを抽出したとき、出力がポジティブとなる確率がサブグループ間で等しい
test fairness	各グループにおいて確信度が同一の入力を抽出したとき、そのラベルがポジティブである確率がグループ間で等しい
well calibration	test fairness に加え、ラベルがポジティブとなる確率が確信度と等しい
balance for positive class	ラベルがポジティブである出力の確信度がグループ間で等しい
balance for negative class	ラベルがネガティブである出力の確信度がグループ間で等しい

資料：Agarwal and Mishra [2021]、Caton and Haas [2020]、Mehrabi *et al.* [2021]、Verma and Rubin [2018]

こうした考え方に基づく主な概念として以下が提案されている。

- fairness through unawareness：配慮が必要な属性を訓練データやテスト・データから排除して機械学習モデルを生成する。
- fairness through awareness：属性群が類似している入力のペアに対して、類似した予測・推論結果を出力する。属性群および予測・推論結果の類似度の尺度と閾値は応用事例に応じて決定される。この概念は individual fairness とも呼ばれる。この特殊ケースとして、配慮が必要な属性以外の属性が同一である入力のペアに対して、同一の予測・推論結果を出力するという概念は causal discrimination と呼ばれることがある。

(3) 属性・出力間の関係性に基づく概念

この分類の概念は、訓練データ等に用いられる属性と機械学習利用システムの出力との間の関係をグラフ等によって表現し、両者の関係性の有無（影響の有無）に着目して定義されるものである。配慮が必要な属性と出力との関係性をグラフ上断ち切ることを目指している。例えば、counterfactual fairness、no proxy discrimination、no unresolved discrimination が挙げられる。

- counterfactual fairness：属性間および属性・出力間の関係を示すグラフにおいて、配慮が必要な属性が出力に影響を与える経路が存在しない。
- no proxy discrimination：同グラフにおいて、配慮が必要な属性が他の属性（proxy attribute）を介して出力に影響を与える経路が存在しない。
- no unresolved discrimination：同グラフにおいて、配慮が必要な属性が、差別（discrimination）につながらない経路を除き、出力に影響を与えない（差別につながらない経路であれば影響を与えてもよい）。