AIシステムにおける透明性と信頼の構築

エンタープライズに求められる「透明性」と「公平性」を担保する「IBM Watson OpenScale」

エンタープライズ・ビジネスにおいてAI導入が進むにつれて、AIというブラックボックスの中から導き出された結論によって引き起こされる倫理的な課題や、その判定結果がもたらす社会的な影響が懸念されています。信頼に足りうるAIでなければ、ビジネスで活用することはできません。この課題を克服するためにAIに求められるのが、導き出された結論に対して「透明性」と「公平性」を担保したガバナンスの確立です。本稿では、AIのブラックボックス化による課題と、その課題を軽減する「IBM Cloud」上の新しいサービス「IBM Watson OpenScale」について解説します。

1. AIのブラックボックス化による問題

多くの企業がAIの可能性を認識し始め、ビジネスへの導入検討や実証実験に乗り出しています。

従来の機械学習では、学習のために着目すべき「特徴量」を人間が指定する必要がありましたが、今日のAIの主流になっているディープ・ラーニングでは、ニューラル・ネットワークと呼ばれる人間の神経構造を模したアルゴリズムを発展させることで、特徴量を自律的に見つけ出せるようになりました。しかし内部ロジックで明確なルールを定義しているわけではないため、膨大な学習データを用いて学習を重ねる中でモデルに組み込まれていった判定因子（ノード）間の複雑な相関関係を人間が見て理解することは容易ではありません。ディープ・ラーニングは精度の高い予測を導き出せる反面、どのような根拠で判断に至ったかを論理的に説明できず、判断基準が「ブラックボックス化」するリスクが指摘されています。

このブラックボックス化により、以下に示す3つの潜在的な問題が発生する可能性があります。

- 透明性の欠如は、ユーザーに結果の妥当性と正確さについて不信感を抱かせ、AIの本質的な価値を損ないます。例えば、医療診断で、AIからいきなり「心臓を手術すべき」と判断されたら、それに素直に応じることができるでしょうか。人間の医師であれば、「あなたの心臓をCTスキャンで検査したところ血栓があり、このままでは心筋梗塞に至る危険性が高い」といった理由をきちんと説明するはずです。AIが導き出した結果も同じで、なぜその結果が導き出されたのか、そのプロセスが明確に説明されなければ信頼することはできません。
- AIにより導き出された結果が人種や性別、地域、年齢などの属性によって偏りがあっても気づかず、社会問題に発展してしまうかもしれません。実際に起きた事例として、ある米国企業のAIアルゴリズムを利用した人事採用システムで、採用に性別による偏りがあり、システムの利用を停止したことが報道されました［1］。このシステムは、過去10年間の就職志願者の履歴書を学習データとして訓練したAIモデルによって志願者の自動ランク分けを行っていましたが、過去の学習データの大半が男性志願者からの履歴書であったため、男性志願者をより優遇する傾向があることが判明しました。後に、判定がより公平になるように改変をしたものの、あらゆるケースにおいて公平性を保証することは難しいと判断され最終的に使用停止になりました。
- AIの思考プロセスが説明できない場合、異常な結果を導き出した際のトラブル・シューティングは困難であるか不可能です。これは、透明性が欠如していると、壊れたものを修正できなくなることを意味しています。

また、企業にとってAIの透明性や公平性は法規制順

守のためにも必要で、AI導入の大きな壁になっています[2]。2016年頃からアカデミアを中心に「Explainable AI」やAIの解釈性に関する研究が急激に増えたのも、この社会的要請からくるものです。

- 一般データ保護規則（GDPR）では「自動化された決定に対抗する権利」として、AIシステムが特定の判断に至った理由を説明してもらうことができる権利が明記されました[3]。
- 内閣府の「人間中心のAI社会原則」の7つの原則の一つに「公平性、説明責任及び透明性の原則」があります。「AIの利用によって人々が不当な差別・扱いを受けることのないように、透明性・公平性のある意思決定と結果に対する説明責任が確保される必要がある」とされています[4]。
- 米国消費者金融保護局が管理する連邦公正信用報告法（FCRA）では、銀行が審査を却下する場合、その理由を具体的に申請者に提供する必要があるとしています。

多くの企業で期待されているビジネス上の広範な「意思決定支援」領域におけるAI活用を考えるときには、このAIのブラックボックス化という課題をクリアしなければなりません。

2. IBMによるAIの信頼性と透明性への取り組み

IBMでは、2017年にコーポレート・リードネスにおいて「AIの信頼性と透明性」を公開しました[5]。2018年にはIBM Researchによる今後5年間に起き得る5つのイノベーションを予測して公開する「5 in 5」でAIにおけるバイアス問題を取り上げ[6]、AI倫理のためのガイドを発表しました[7]。このガイドでは、5つの信頼と透明性に関する原則（**表1**）が述べられており、倫理的な意思決定が単なる技術的な問題解決ではなく、人間生活・関係をより豊かにするための仕組みの一つがAIであると位置付けています。現在進行系でもあるこのガイドの策定は、複数の専門分野からの参画だけではなく、AI倫理に関心がある一般の参画に基づいており、よりよい倫理に関する取り組みと枠組みが構成されると期待されています。この5つの重要分野で定義されたExplainability（説明可能性）とFairness（公平性）を具現化するためのサービスがIBM Watson OpenScale（以下、Watson OpenScale）[8]です。

表1. ガイドで定義された5つの重要分野
（Everyday Ethics for Artificial Intelligenceより）

Accountability（説明責任）	AIの設計者と開発者は、AIの設計、開発、意思決定プロセス、結果に対して熟慮する責任を負います。
Value Alignment（価値観の一致）	AIの設計は、対象とするユーザー・グループが有する規範や価値観を考慮して行うべきです。
Explainability（説明可能性）	AIの決定プロセスが人間にも容易に認知、感知、理解ができるように、AIを設計すべきです。
User Data Rights（ユーザー・データの権利）	ユーザー・データを保護し、アクセスや利用に関するユーザーの権利を保持できるように、AIを設計すべきです。
Fairness（公平性）	バイアス（偏り）を最小限に抑え、誰もが参加できる社会を後押しするように、AIを設計すべきです。

3. IBM Watson OpenScaleの概要

Watson OpenScaleは、全社規模のAIの自動化と運用を行うIBM Cloud上のサービスです。企業内にデプロイしている各AIモデルをモニターし、大きく以下の4つの機能を提供します。

- **Fairness（公平性）：**本番環境におけるAIモデルの予測バイアス（偏り）の検知
- **De-Bias（バイアスの軽減）：**再学習なしに、自動的にAIモデルの予測バイアスを軽減
- **Explainability（説明可能性）：**AIモデルの予測根拠の説明
- **Accuracy（精度）：**AIモデルの予測性能劣化の検知

どのようなAIモデルでも、ある程度のバイアスがある可能性があります。モデルは与えられた学習データ以上のことはできません。学習に使用したデータセットは実社会のデータを100%表すことができないため、新しく学習されたモデルが本番環境でうまく機能しない可能性があります。さらに、ほとんどのデータドメインは絶えず進化しており、モデルの精度は時間の経過とともに変動する傾向にあります。重要なのは、AIにより意思決定が行われたタイミングでその説明とAIモデル内のバイアス検知を可能にし、潜在的に不公平な結果が起こり得ることを捕捉し、可視化することです。運用ライフサイクルを通してAIモデルの精度、パフォーマンス、および公平性を監視し、基幹業務ユーザーが結果の根拠を理解するのに役立つ分析を提供することができれば、AIのブラックボックス化解消の大きな一歩となります。

Watson OpenScaleのダッシュボードには企業内で運用している複数の学習モデルを一覧表示でき、それぞれのモデルごとに「精度」と「公平性」のスコアが定量的な数値（%）で表示されます。またスコアが一定の基準以下だった場合には、色を変えてハイライト表示されます。このように可視化された環境のもとで、学習モデルに悪影響を及ぼしているバイアスを自動検知して軽減し、公平な結果を生成するように修正を施すことが可能となります。精度は、過去にモデルに入力されたデータとその正解を含むフィードバック・データ群を用いて算出される予測精度の値であるため、バイアスを軽減すると反対に精度が落ちることがあります。「精度」と「公平性」のトレードオフを評価し、組織のKPIと関連付け、ビジネス状況の変化を考慮して、AIモデルを改善することを可能にするのがWatson OpenScaleです。

Watson OpenScaleは、「IBM Watson Studio」や「IBM Watson Machine Learning」などのAIモデルを構築・実行するためのIBMツールとシームレスな統合が可能です。加えて、他のベンダーのモデル開発環境やオープンソース・ツールと簡単に連携できるオープン・プラットフォームとして設計されており、TensorFlow、Keras、SparkML、Seldon、AWS SageMaker、AzureMLなど多様な機械学習フレームワーク、ライブラリー、AI開発環境に対応しています。パブリック、プライベート、オンプレミスのどの環境とも統合でき、AIのデプロイに関して柔軟で、オープンな選択肢を維持することが可能です。

4. AIによる判断の公平性

そもそもAIによる判断に求められる「公平性」とは何でしょう。技術的な世界と法制や規制の世界では、公平性の意味が異なります。技術的な世界では、「統計学的なバイアス（偏り）」を意味し、法制などでは「不平等」「不公平」という社会正義や福祉という範疇の判断が求められる風潮があります。Watson OpenScaleでは公平性を前者の技術的な世界での用語として使用しています。

では、統計学的なバイアスとはどういうものでしょうか。その意味を理解する上で必要ないくつかの重要な概念を、保険金請求を承認するか拒否するかについて決定を下すAIシステムの例で考えてみます。

● フェアネス属性（Fairness Attribute）

バイアスまたは公平性は通常、法律やポリシーで定められた性別、人種、年齢といった保護対象になっている属性（Protected Class[9]）を使用して測定されますが、

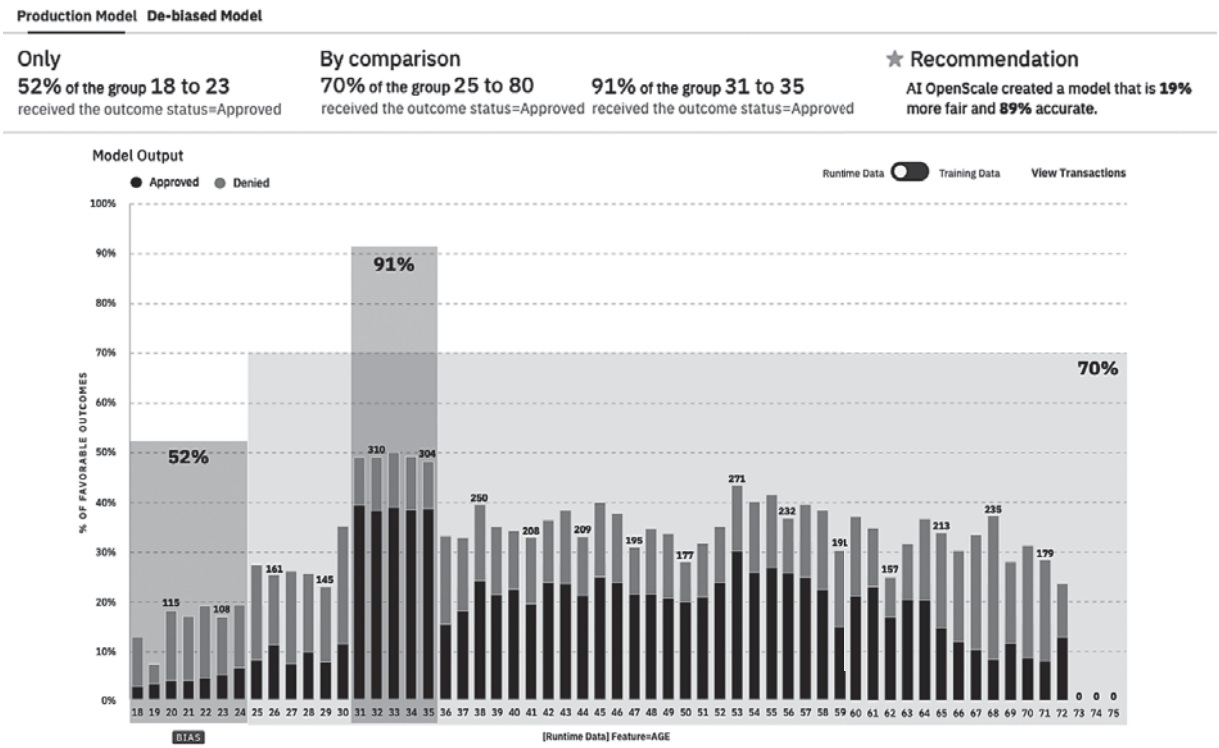

図1. 保険契約業務での承認状況の提示。18〜24歳に対する契約承認に偏りがあることを示唆している。

Watson OpenScaleでは、保険の加入年数といった値を使用して測定することもできます。

● 監視対象グループ/参照グループ (Monitored group/Reference group)

監視対象グループは、バイアスを測定したいフェアネス属性の値です。フェアネス属性のもう一方の値は参照グループと呼ばれます。フェアネス属性を「性別」として女性に対するバイアスを測定しようした場合、監視対象グループは「女性」となり、参照グループは「男性」になります。

● 好ましい/好ましくない結果 (Favorable/UnFavorable Outcome)

バイアス検出で重要な概念は、モデルの「好ましい結果」と「好ましくない結果」です。例えば、承認された請求は「好ましい」結果と見なすことができ、拒否された請求は「好ましくない結果」と見なすことができます。

● 差別的効果 (Disparate Impact) [10]

これはバイアスを測定するために使用され、「監視グループに好ましい結果が発生した割合」/「参照グループに好ましい結果が発生した割合」(以下、Impact Ratio)として計算されます。例えば、男性による請求の80%が承認されているのに対し、女性による請求の60%だけが承認されている場合、Impact Ratioの値は60/80=0.75になります。通常、この値が0.8を下回っているとバイアスが存在すると判断されます。

● バイアスのしきい値

Impact Ratioの値が、バイアス有無判定のしきい値になります。米国の過去の判例やガイドラインでは、Impact Ratioの値が0.8であれば公平で、差異は許容範囲であると見なされます(これを4/5ルール、または80%ルールと呼びます[11])。

4-1. バイアスの検知

Watson OpenScaleはAIモデルの実行時バイアスを検出することができ、モデルに送信されたデータとモデルの予測結果をモニターします(このデータはペイロード・データと呼ばれます)。ペイロード・データを内部のデータベースに保存していき、1時間ごとに直近の時間帯のDisparate Impactの評価を行います。**図1**の例ではフェアネス属性は「年齢(Age)」で、18〜24歳の人たちの好ましい結果の割合は52%であるのに対し、25〜75歳の人たち(参照グループ)の好ましい結果の割合は70%です。従って、Impact Ratioは0.52/0.7=0.74と算出でき、0.8のしきい値を下回っているため、このモデルにはバイアスが存在していると見なされます。

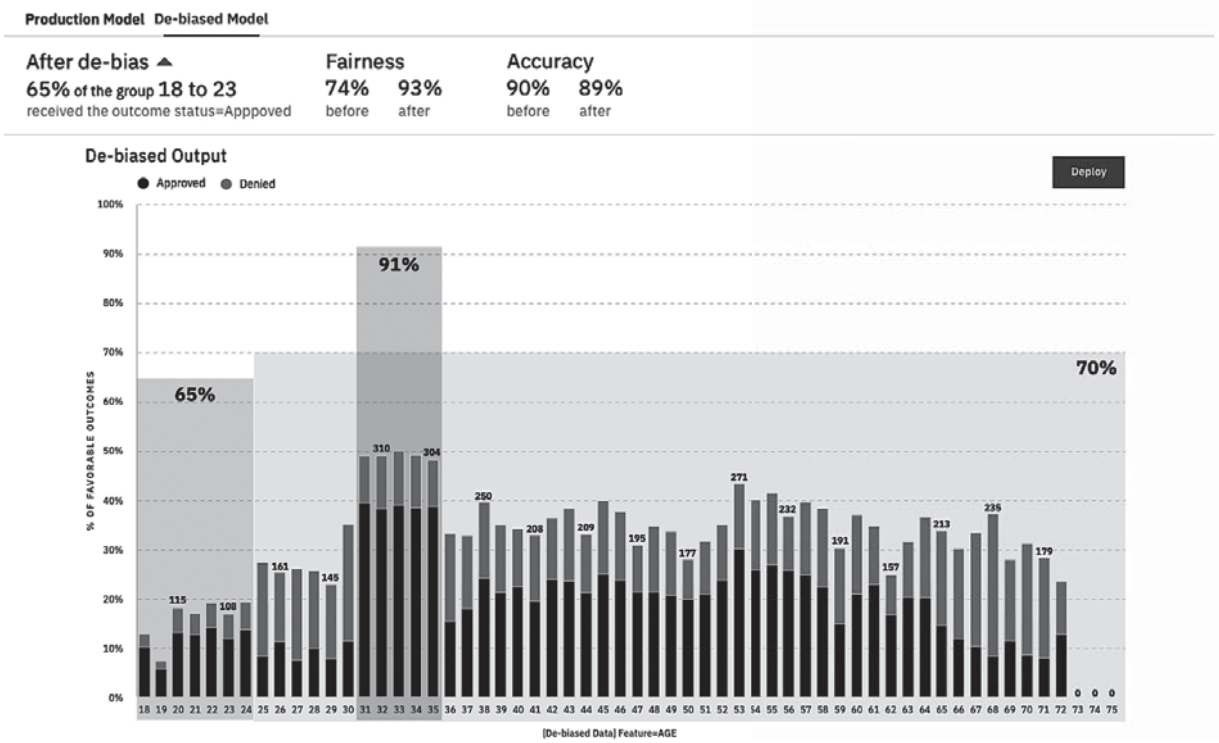

図2. バイアスの軽減例:18〜24歳の好ましい結果の割合が52%から65%まで上がっている。

4-2. バイアスの軽減

Watson OpenScaleには、バイアスを検知する機能に加え、再学習せずに自動的にバイアスを軽減する機能を備えています。Watson OpenScaleは、1時間おきにこれまで蓄積してきたペイロード・データに対してIBM Research独自のバイアス軽減アルゴリズム[12]を実行します（Passive De-Biasing）。アルゴリズム適用の効果はダッシュボード上で確認できます（**図2**）。また、アプリケーションに別のRESTエンドポイントを提供し、モデルへのスコアリング要求時に自動的にバイアス軽減を行う（Active De-Biasing）ことも可能で、通常は、Passive De-Biasingで効果を確認したあとで、Active De-Biasingに切り替えることになります。

5. AIによる判断の可視化と説明可能性

Watson OpenScaleは、次の2種類の説明を提供することでこの問題に対処します。

● LIMEベースの説明

LIME（Local Interpretable Model-agnostic Explanations）アルゴリズムを使用して、予測結果にプラスまたはマイナスに働く特徴量を見つけます。例えば、保険金請求が拒否（または承認）された理由を説明します。

● Contrastive Explanations

IBM Research独自のテクノロジー[13]で、モデル全体の特性や振る舞いの説明を行います。例えば、このモデルはどのような入力値だと拒否する傾向にあるのか、どういう入力値だと承認されるといった情報を提供します。

LIMEは、モデルへ入力する特徴量を外部からさまざまに変化させてみて（これを摂動 - Perturbationと呼びます。例えば「性別」フィールドの「女性」という値を「男性」に変更してみる等）、予測の結果が変化するかどうかを調査する手法です。これは、外部から与える特徴量の変化と予測結果の変化の関係に着目した説明性を提示する手法で、説明対象のAIモデルの内部実装に依存せずに済むという利点があります。Watson OpenScaleでは、説明対象のモデル実装に依存しないように独自改良したアルゴリズムを利用しています。これにより、IBMが提供する製品・サービスのAIだけでなく、他社がサポートしているAIにも対応でき、説明が困難なアンサンブル・モデルやディープ・ラーニング（文章・画像）のような「ブラックボックス」モデルに関しても、根拠や主要因を示すことができるようになります。

これらの概念を理解するために、保険金請求を承認するか拒否するかを決定するモデルを考えてみます。**図3**は、

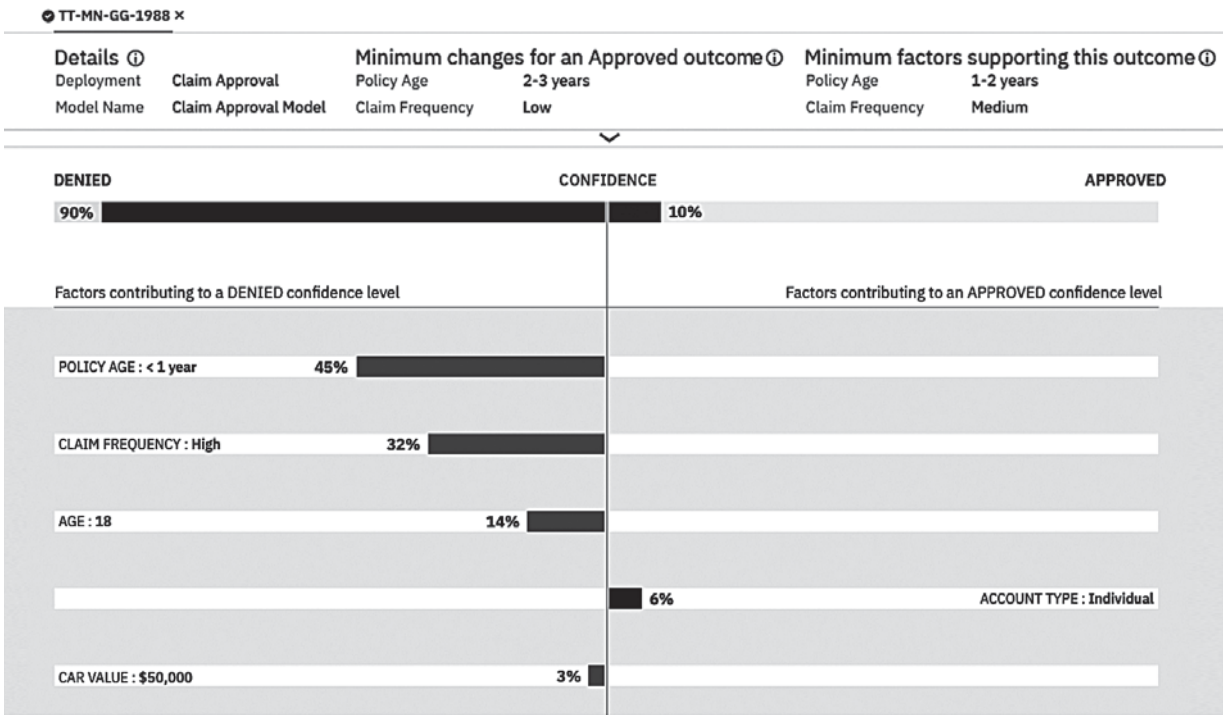

図3. 保険金請求が却下された理由の説明

Watson OpenScaleを使用して、「この保険金請求が拒否されたのはなぜですか？」という顧客または監査部門からの問い合わせに対して、その答えを提示したものです。

まずLIMEベースの説明を示す**図3**の下部に注目してみます。LIMEベースの説明は、予測にプラスにもマイナスにも寄与した一連の特徴量を示しています。この例では、①保険契約年数が1年未満、②保険金請求頻度が多い、③年齢が18歳、という3つの特徴量が「却下」と判断する際の根拠になっていることを示しています。一方、アカウントの種類が「Individual(個人)」であるという事実は、請求が承認されるべきである可能性を示しています。

次に、Contrastive Explanationsを示す**図3**の上部に注目してみます。「Minimum changes for an Approved outcome」は、モデルによる予測が「拒否」から「承認」に変わるために必要な特徴量の最小の変化を示しています。言い換えれば保険加入年数が2〜3年で、請求頻度が低い場合、モデルはその請求が承認されると予測していたことを示しています。「Minimum factors supporting this outcome」は、予測結果に変化をもたらさなかった特徴値の最大の変化を示します。この例では、「契約年数」が1〜2年で「申し立ての頻度」が「中」であったとしても、モデルの予測は「拒否」のままであることを示しています。このContrastive Explanationsにより、顧客に対して却下理由だけでなく、「承認されるためにはどうすればよいか」という、より建設的な説明が可能になります。

6. おわりに

本稿では、AIをビジネスで活用するために必要となる公平性と説明可能性を提供するサービス、Watson OpenScaleを紹介しました。AIシステムの性能をリアルタイムにモニターすることが可能となり、AIシステムのサービス品質が把握できるようになります。内閣府の指針に示されているように、今後、AIの公平性、説明可能性、透明性の実現は、AIが適切に社会に根付く上で重要な原則として位置付けられています。企業において全社的にAIをモニタリングしていく機能は、必須のものになっていくでしょう。Watson OpenScaleにより、エンタープライズでのAIの利活用がより推進されることを期待しています。

[参考文献]

[1] MIT Technology Review:「女子大卒は減点」アマゾンのAI採用、男性優遇判明で廃止, MIT Technology Review, https://www.technologyreview.jp/nl/amazon-ditched-ai-recruitment-software-because-it-was-biased-against-women/

[2] IBM Institute for Business Value:「エンタープライズAIへのシフト」, https://www.ibm.com/services/jp-ja/studies/thoughtleadership/

[3] 第17回:AI・ロボット法の文脈における欧州一般データ保護規則(GDPR)の「自動化された決定」に対抗する権利, http://www.kbd-personalinfo.com/entry/2018/01/12/140115

[4] 人間中心の AI 社会原則(案): https://www8.cao.go.jp/cstp/ai_gensoku.pdf

[5] IBM 2017 Corporate Responsibility Report: Putting smart to work for our company and the world, https://www.ibm.com/ibm/responsibility/2017/

[6] IBM 5 in 5: AI bias will explode. But only the unbiased AI will survive, https://www.research.ibm.com/5-in-5/ai-and-bias/

[7] IBM Everyday Ethics for Artificial Intelligence:https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

[8] Watson OpenScale:信頼できるAIの採用を加速させるオープン・プラットフォーム, https://www.ibm.com/watson/jp-ja/ai-openscale/

[9] Protected Group:https://en.wikipedia.org/wiki/Protected_group

[10] WIKIPEDIA, Disparate impact, https://en.m.wikipedia.org/wiki/Disparate_impact

[11] The Four/Fifths Rule:https://www.mbaskool.com/business-concepts/human-resources-hr-terms/13006-45ths-rule.html

[12] BIAS MITIGATION POST-PROCESSING FOR INDIVIDUAL AND GROUP FAIRNESS, https://arxiv.org/abs/1812.06135

[13] Model Agnostic Contrastive Explanation for Machine Learning Classification Models, https://www.ibm.com/downloads/cas/0ZRZNR8E

日本アイ・ビー・エム株式会社
クラウド＆コグニティブ・ソフトウェア
テクニカル・リード

土屋 敦
Atsushi Tsuchiya

2008年日本IBM入社。組み込みデータベース、ストリーム処理ミドルウェアを担当し、現在はIoTとData&AI領域におけるソリューション・アーキテクト担当として活動中。

日本アイ・ビー・エム株式会社
パートナー・アライアンス事業本部
ソリューション・アーキテクト

石田 剛
Tsuyoshi Ishida

1984年日本IBM入社。都銀アカウントSEを皮切りに、さまざまなSIプロジェクトに従事。現在はパートナー担当のソリューション・アーキテクトとして活動中。

日本アイ・ビー・エム株式会社
研究開発
シニア・ソフトウェア・エンジニア
IBMアカデミー・オブ・テクノロジー・メンバー
IBMマスター・インベンター

大谷 宗孝
Munetaka Ohtani

日本IBM入社以来、一貫してソフトウェア製品開発に従事。IBM Analytics製品の基幹ソフトウェアInformation Server DataStageや、ヘルスケア領域でのクラウドの開発を行う。現在はお客様とIBM研究開発部門の「共創」を支援。