(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2024-520312
(P2024-520312A)
(43)公表日　令和6年5月24日(2024.5.24)

(51)国際特許分類　FI
G06N 3/08 (2023.01)　G06N 3/08

審査請求　未請求　予備審査請求　未請求　（全30頁）

(21)出願番号　特願2023-570346(P2023-570346)
(86)(22)出願日　令和4年6月1日(2022.6.1)
(85)翻訳文提出日　令和5年11月14日(2023.11.14)
(86)国際出願番号　PCT/IB2022/055104
(87)国際公開番号　WO2022/259089
(87)国際公開日　令和4年12月15日(2022.12.15)
(31)優先権主張番号　17/303,732
(32)優先日　令和3年6月7日(2021.6.7)
(33)優先権主張国・地域又は機関
米国(US)
(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,
最終頁に続く

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ ＢＵＳＩＮＥＳＳ ＭＡＣＨＩＮＥＳ ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国１０５０４ ニューヨーク州 アーモンク ニュー オーチャード ロード
Ｎｅｗ Ｏｒｃｈａｒｄ Ｒｏａｄ， Ａｒｍｏｎｋ， Ｎｅｗ Ｙｏｒｋ １０５０４， Ｕｎｉｔｅｄ Ｓｔａｔｅｓ ｏｆ Ａｍｅｒｉｃａ
(74)代理人　100112690
弁理士　太佐 種一
最終頁に続く

(54)【発明の名称】　人工知能モジュール訓練中のバイアス低減

(57)【要約】
本明細書において、１つまたは複数の選ばれた変数を含む入力データ・セットを受け取ることに応じて分析結果を提供するように訓練される、調節可能なパラメータをもつ人工知能モデルを訓練する方法が開示される。この方法は、訓練分析結果と対になる訓練入力データの多数のグループを含む訓練データ・セットを受け取ることと、訓練入力データの多数のグループを人工知能モデルに入力することに応じて、人工知能モデルから試行分析結果を受け取ることと、試行分析結果と訓練分析結果との間の比較を記述する正確度メトリックを計算することと、１つまたは複数の選ばれた変数を試行分析結果と比較することによって公平性スコア・メトリックを計算することと、公平性スコア・メトリックおよび正確度メトリックから組み合わせメトリックを計算することと、少なくとも組み合わせメトリックを受け取る訓練アルゴリズムを使用して、調節可能なパラメータを修正することとを含む。

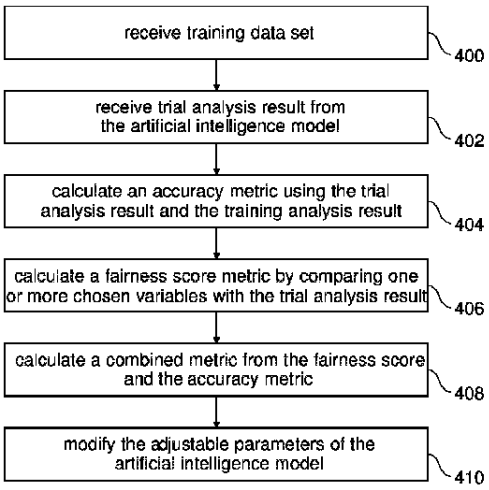


Fig. 4

【特許請求の範囲】
【請求項１】
人工知能モデルを訓練する方法であって、前記人工知能モデルが、調節可能なパラメータを有し、前記人工知能モデルが、入力データ・セットを受け取ることに応じて、分析結果を提供するように訓練され、前記入力データ・セットが、１つまたは複数の選ばれた変数を含み、前記方法が、
前記人工知能モデルを訓練するための訓練データ・セットを受け取ることであり、前記訓練データ・セットが、訓練分析結果と対になる訓練入力データの多数のグループを含む、前記受け取ることと、
前記訓練入力データの多数のグループを前記入力データ・セットとして前記人工知能モデルに入力することに応じて、前記人工知能モデルから試行分析結果を受け取ることと、
前記試行分析結果と前記訓練分析結果との間の比較を記述する正確度メトリックを計算することと、
前記１つまたは複数の選ばれた変数を前記試行分析結果と比較することによって公平性スコア・メトリックを計算することと、
前記公平性スコア・メトリックおよび前記正確度メトリックから組み合わせメトリックを計算することと、
少なくとも前記組み合わせメトリックを入力として受け取る訓練アルゴリズムを使用して、前記人工知能モデルの前記調節可能なパラメータを修正することと
を含む、方法。
【請求項２】
前記方法が、
多数の訓練された人工知能モデルを受け取ることであり、前記多数の訓練された人工知能モデルが前記人工知能モデルからなる、前記受け取ることと、
前記多数の人工知能モデルをテストするためのテスト・データ・セットを受け取ることであり、前記テスト・データ・セットが、テスト分析結果と対になるテスト入力データの多数のグループを含む、前記受け取ることと、
前記テスト入力データの多数のグループを前記入力データ・セットとして入力することに応じて、前記多数の人工知能モデルの各々から緩和分析結果を受け取ることと、
前記多数の訓練された人工知能モデルの各々に対する前記緩和分析結果と、前記テスト分析結果との間の比較を記述する、前記多数の訓練された人工知能モデルの各々に対する正確度スコアを計算することと、
前記１つまたは複数の選ばれた変数を前記試行分析結果と比較することによって、前記多数の訓練された人工知能モデルの各々に対する公平性評価メトリックを計算することと、
前記多数の訓練された人工知能モデルの各々に対して前記公平性評価メトリックと前記正確度スコアとを組み合わせることによって、前記多数の訓練された人工知能モデルの各々に対する公平性重み付きランキングを計算することと
によって、前記多数の訓練された人工知能モデルの各々に対する前記公平性重み付きランキングを提供することをさらに含む、請求項１に記載の方法。
【請求項３】
前記公平性評価メトリックが、前記１つまたは複数の選ばれた変数の１つまたは複数の選ばれた値と、前記試行分析結果との間の相関を記述する、請求項２に記載の方法。
【請求項４】
前記多数の訓練された人工知能モデルが、異なるタイプのものである、請求項２または３に記載の方法。
【請求項５】
前記多数の訓練された人工知能モデルの各々が、独立して、以下のもの、すなわち、ニューラル・ネットワーク、分類器ニューラル・ネットワーク、畳み込みニューラル・ネットワーク、ベイジアン・ニューラル・ネットワーク、ベイジアン・ネットワーク、ベイズ

・ネットワーク、単純ベイズ分類器、信念ネットワーク、または決定ネットワーク、決定木、サポート・ベクトル機械、回帰分析、および遺伝的アルゴリズムのうちの任意の１つである、先行する請求項２ないし４のいずれか一項に記載の方法。
【請求項６】
前記公平性重み付きランキングが、以下のもの、すなわち、前記公平性評価メトリックと前記正確度スコアの最小２乗組み合わせ、前記公平性評価メトリックと前記正確度スコアの重み付き最小２乗組み合わせ、前記公平性評価メトリックと前記正確度スコアの線形組み合わせ、前記公平性評価メトリックと前記正確度スコアの重み付き組み合わせ、および前記公平性評価メトリックと前記正確度スコアの多項式組み合わせのうちの任意の１つを含む、請求項１ないし５のいずれか一項に記載の方法。
【請求項７】
前記組み合わせメトリックが、前記正確度スコアにスケーリング係数を乗算し、それを所定の冪乗したものであり、前記スケーリング係数が、前記公平性評価メトリックの関数である、請求項２ないし５のいずれか一項に記載の方法。
【請求項８】
前記スケーリング係数が、前記公平性評価メトリックの逆数である、請求項７に記載の方法。
【請求項９】
前記公平性スコア・メトリックが、前記１つまたは複数の選ばれた変数の１つまたは複数の選ばれた値と、前記試行分析結果との間の相関を記述する、請求項１ないし８のいずれか一項に記載の方法。
【請求項１０】
前記組み合わせメトリックが、以下のもの、すなわち、前記公平性スコア・メトリックとテスト・メトリックの最小２乗組み合わせ、前記公平性スコア・メトリックと前記テスト・メトリックの重み付き最小２乗組み合わせ、前記公平性スコア・メトリックと前記テスト・メトリックの線形組み合わせ、前記公平性スコア・メトリックと前記テスト・メトリックの重み付き組み合わせ、および前記公平性スコア・メトリックと前記テスト・メトリックの多項式組み合わせのうちの任意の１つを含む、請求項１ないし９のいずれか一項に記載の方法。
【請求項１１】
前記組み合わせメトリックが、以下のもの、すなわち、前記公平性スコア・メトリックに対する制約、前記テスト・メトリックに対する制約、前記公平性スコア・メトリックに対する最大許容値、および前記テスト・メトリックに対する最大許容値のうちの任意の１つを含む、請求項９または１０に記載の方法。
【請求項１２】
前記人工知能モデルが、以下のもの、すなわち、ニューラル・ネットワーク、分類器ニューラル・ネットワーク、畳み込みニューラル・ネットワーク、ベイジアン・ニューラル・ネットワーク、ベイジアン・ネットワーク、ベイズ・ネットワーク、単純ベイズ分類器、信念ネットワーク、または決定ネットワーク、決定木、サポート・ベクトル機械、回帰分析、および遺伝的アルゴリズムのうちの任意の１つである、請求項１ないし１１のいずれか一項に記載の方法。
【請求項１３】
前記人工知能モデルが畳み込みニューラル・ネットワークであり、前記訓練アルゴリズムが深層学習アルゴリズムである、請求項１ないし１２のいずれか一項に記載の方法。
【請求項１４】
コンピュータ可読プログラム・コードが具現化されたコンピュータ可読ストレージ媒体を含むコンピュータ・プログラム製品であって、前記コンピュータ可読のプログラム・コードが、請求項１ないし１３に記載の前記方法を実施するように構成される、コンピュータ・プログラム製品。
【請求項１５】

コンピュータ・システムを制御するように構成されたプロセッサと、
機械実行命令を格納するメモリであって、前記命令の実行により、前記プロセッサが、
人工知能モデルを訓練するための訓練データ・セットを受け取ることであり、前記人工知能モデルが、調節可能なパラメータを有し、前記人工知能モデルが、入力データ・セットを受け取ることに応じて、分析結果を提供するように訓練され、前記入力データ・セットが、１つまたは複数の選ばれた変数を含み、前記訓練データ・セットが、訓練分析結果と対になる訓練入力データの多数のグループを含む、前記受け取ることと、
前記訓練入力データの多数のグループを前記入力データ・セットとして前記人工知能モデルに入力することに応じて、前記人工知能モデルから試行分析結果を受け取ることと、
前記試行分析結果と前記訓練分析結果との間の比較を記述する正確度メトリックを計算することと、
前記１つまたは複数の選ばれた変数を前記試行分析結果と比較することによって計算される公平性スコア・メトリックを計算することと、
前記公平性スコア・メトリックおよび前記正確度メトリックから組み合わせメトリックを計算することと、
少なくとも前記組み合わせメトリックを入力として受け取る訓練アルゴリズムを使用して、前記人工知能モデルの前記調節可能なパラメータを修正することと
を行う、メモリと
を含むコンピュータ・システム。
【請求項１６】
前記命令の実行により、さらに、前記プロセッサが、
多数の訓練された人工知能モデルを受け取ることであり、前記多数の訓練された人工知能モデルが前記人工知能モデルからなる、前記受け取ることと、
前記多数の人工知能モデルをテストするためのテスト・データ・セットを受け取ることであり、前記テスト・データ・セットが、テスト分析結果と対になるテスト入力データの多数のグループを含む、前記受け取ることと、
前記テスト入力データの多数のグループを前記入力データ・セットとして入力することに応じて、前記多数の人工知能モデルの各々から緩和分析結果を受け取ることと、
前記多数の訓練された人工知能モデルの各々に対する前記緩和分析結果と、前記テスト分析結果との間の比較を記述する、前記多数の訓練された人工知能モデルの各々に対する正確度スコアを計算することと、
前記１つまたは複数の選ばれた変数を前記試行分析結果と比較することによって、前記多数の訓練された人工知能モデルの各々に対する公平性評価メトリックを計算することと、
前記多数の訓練された人工知能モデルの各々に対して前記公平性評価メトリックと前記正確度スコアとを組み合わせることによって、前記多数の訓練された人工知能モデルの各々に対する公平性重み付きランキングを計算することと
を行う、請求項１５に記載のコンピュータ・システム。
【請求項１７】
前記人工知能モデルが、以下のもの、すなわち、ニューラル・ネットワーク、分類器ニューラル・ネットワーク、畳み込みニューラル・ネットワーク、ベイジアン・ニューラル・ネットワーク、ベイジアン・ネットワーク、ベイズ・ネットワーク、単純ベイズ分類器、信念ネットワーク、または決定ネットワーク、決定木、サポート・ベクトル機械、回帰分析、および遺伝的アルゴリズムのうちの任意の１つである、請求項１５または１６に記載のコンピュータ・システム。
【請求項１８】
前記人工知能モデルが畳み込みニューラル・ネットワークであり、前記訓練アルゴリズムが深層学習アルゴリズムである、請求項１５ないし１７のいずれか一項に記載のコンピュータ・システム。

【請求項１９】
コンピュータ・プログラム製品であって、前記コンピュータ・プログラム製品が、請求項１ないし１３に記載の前記方法に従って訓練された人工知能モデルを格納したコンピュータ可読ストレージ媒体を含む、コンピュータ・プログラム製品。
【請求項２０】
データ処理システムで実行されるアプリケーション・プログラムがアクセスするためのデータを格納するためのメモリであって、請求項１ないし１３に記載の前記方法に従って訓練される人工知能モデルを含む、メモリ。
【請求項２１】
多数の訓練された人工知能モデルの各々に公平性重み付きランキングを提供する方法であって、前記方法が、
前記多数の訓練された人工知能モデルを受け取ることと、
前記多数の人工知能モデルをテストするためのテスト・データ・セットを受け取ることであり、前記テスト・データ・セットが、テスト分析結果と対になるテスト入力データの多数のグループを含む、前記受け取ることと、
前記テスト入力データの多数のグループを入力データ・セットとして入力することに応じて、前記多数の人工知能モデルの各々から緩和分析結果を受け取ることと、
前記多数の訓練された人工知能モデルの各々に対する前記緩和分析結果と、前記テスト分析結果との間の比較を記述する、前記多数の訓練された人工知能モデルの各々に対する正確度スコアを計算することと、
１つまたは複数の選ばれた値を試行分析結果と比較することによって、前記多数の訓練された人工知能モデルの各々に対する公平性評価メトリックを計算することと、
前記多数の訓練された人工知能モデルの各々に対して前記公平性評価メトリックと前記正確度スコアとを組み合わせることによって、前記多数の訓練された人工知能モデルの各々に対する前記公平性重み付きランキングを計算することと
を含む、方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、人工知能モデルの訓練に関する。
【背景技術】
【０００２】
人工知能（ＡＩ）モジュールおよびアルゴリズムの自動訓練は、非常に普及しており、それらを訓練するために必要とされる人間の労力の低減を可能にする。しかしながら、好結果の訓練は、現在、人工知能モジュールが系統的なバイアスまたは偏見を含む結果を生成しないように、訓練を非常に注意深く実行することに依拠する。現在、訓練データが系統的なバイアスを含む場合、訓練された人工知能モジュールもそうなることになる。
【発明の概要】
【０００３】
１つの態様では、本発明は、人工知能モデルを訓練する方法を提供する。人工知能モデルは、調節可能なパラメータを有する。調節可能なパラメータは、人工知能モデルの実行および動作に影響を与える。それゆえに、人工知能モデルは、調節可能なパラメータを修正または調節することによって訓練することができる。人工知能モデルは、入力データ・セットを受け取ることに応じて、分析結果を提供するように訓練される。入力データ・セットは、１つまたは複数の選ばれた変数を含む。
【０００４】
この方法は、人工知能モデルを訓練するための訓練データ・セットを受け取ることを含む。訓練データ・セットは、訓練分析結果と対になる訓練入力データの多数のグループを含む。訓練入力データは、人工知能モデルへの入力としての試行として使用されるデータとすることができる。次いで、人工知能モデルの出力は、訓練分析結果と比較することが

できる。この方法は、訓練入力データの多数のグループを入力データとして人工知能モデルに入力することに応じて、人工知能モデルから試行分析結果を受け取ることをさらに含む。このステップにおいて、訓練入力データが人工知能モデルに入力され、それに応じて、試行分析結果が受け取られる。この方法は、前記試行分析結果と前記訓練分析結果との間の比較を記述する正確度（accuracy）メトリックを計算することをさらに含む。人工知能モデルから得られる結果である試行分析結果は、訓練分析結果と比較され、正確度メトリックは、試行分析結果が訓練分析結果にどれだけ近いかまたは正確であるかを評価する尺度または値を提供する。

【０００５】

この方法は、１つまたは複数の選ばれた変数を試行分析結果と比較することによって公平性スコア・メトリックを計算することをさらに含む。人工知能の公平性尺度または公平性スコアは、特定の変数、またはこの場合には１つまたは複数の選ばれた変数が、人工知能モデルの出力にどれだけ影響を与えるかの尺度を指す。

【０００６】

この方法は、公平性スコア・メトリックおよび正確度メトリックから組み合わせメトリックを計算することをさらに含む。この方法は、少なくとも前記組み合わせメトリックを入力として受け取る訓練アルゴリズムを使用して、人工知能モデルの調節可能なパラメータを修正することをさらに含む。

【０００７】

本発明のさらなる態様によれば、本発明は、プロセッサと、機械実行可能命令を格納するメモリとを含むコンピュータ・システムを提供する。機械実行可能命令の実行により、プロセッサは、一実施形態による方法を実施する。

【０００８】

本発明のさらなる態様によれば、本発明は、コンピュータ可読プログラム・コードが具現化されたコンピュータ可読ストレージ媒体を含むコンピュータ・プログラム製品を提供する。コンピュータ可読プログラム・コードは、一実施形態による方法を実施するように構成される。

【０００９】

本発明のさらなる態様によれば、本発明は、コンピュータ・プログラム製品を提供する。コンピュータ・プログラム製品は、この方法の一実施形態に従って訓練された人工知能モデルを格納したコンピュータ可読ストレージ媒体を含む。

【００１０】

本発明のさらなる態様によれば、本発明は、データ処理システムで実行されるアプリケーション・プログラムがアクセスするためのデータを格納するメモリを提供する。これは、この方法の一実施形態に従って訓練された人工知能モデルを含む。

【００１１】

以下において、本発明の実施形態が、単に例として、図面を参照してより詳細に説明される。

【図面の簡単な説明】

【００１２】

【図１】コンピュータ・システムの一例を示す図である。

【図２】図１のコンピュータ・システムが接続される例示的なコンピューティング環境を示す図である。

【図３】コンピュータ・システムのさらなる例を示す図である。

【図４】図３のコンピュータ・システムを使用する方法を示す流れ図である。

【図５】コンピュータ・システムのさらなる例を示す図である。

【図６】図５のコンピュータ・システムを使用する方法を示す流れ図である。

【発明を実施するための形態】

【００１３】

本発明の様々な実施形態の説明は、例証の目的のために提示されるが、網羅的であるこ

と、または開示される実施形態に限定されることを意図するものではない。説明される実施形態の範囲および思想から逸脱することなく、多くの変更および変形が当業者には明らかであろう。本明細書で使用される用語は、実施形態の原理、実際の適用、もしくは市場で見いだされる技術に対する技術的改善を最も良く説明するために、または本明細書で開示される実施形態を他の当業者が理解できるようにするために選ばれた。

【００１４】

実施形態は、１つまたは複数の選ばれた変数に対して望ましくないバイアスを低減する手段を提供することができるので有益であり得る。これは、例えば、訓練データ・セットが望ましくないバイアスまたは偏見を含むにもかかわらず、低減されたバイアスで人工知能モジュールを訓練することを可能にすることができる。

【００１５】

例えば、人工知能モデルは、機械の保守が実行されるべきかどうかおよびいつ実行されるべきかを評価するように訓練される。人工知能モデルを訓練するために使用されるデータには以前の経験および個人的な好みに起因するバイアスがある場合がある。

【００１６】

通常、人工知能モデルが訓練されるとき、正確度メトリックのみが、調整可能パラメータを評価し、次いで、修正するために使用される。組み合わせメトリックは、正確な結果を提供するための人工知能モデルの必要性と、いわゆる公平な結果を提供することの必要性とをバランスさせる手段を提供することができる。それは、特定の変数における、またはこの場合、１つまたは複数の選ばれた変数における望ましくないバイアスを排除しようと試みることである。

【００１７】

例えば、訓練アルゴリズムの入力としてただ単に正確度メトリックを使用する代わりに、組み合わせメトリックが代わりに使用される。すぐ上で説明したように、これは、１つまたは複数の選ばれた変数における望ましくないバイアスを除去する手段を提供することができる。ニューラル・ネットワークの例では、正確度メトリックは損失関数とすることができる。ニューラル・ネットワークの場合、正確度メトリックの結果の代わりに、組み合わせメトリックを、逆伝播アルゴリズムへの入力として使用することができる。ニューラル・ネットワークでは、組み合わせメトリックは、公平性スコア・メトリックの値を通常の損失関数または従来の損失関数と組み合わせる修正された損失関数になることになる。

【００１８】

別の実施形態では、この方法は、多数の訓練された人工知能モデルを最初に受け取ることによって、多数の訓練された人工知能モデルの各々に対して公平性重み付きランキングを提供することをさらに含む。多数の訓練された人工知能モデルは、人工知能モデルからなる。多数の訓練された人工知能モデルの各々に対する公平性重み付きランキングは、例えば、多数の訓練された人工知能モデルの各々が１つまたは複数の選ばれた変数においてどれだけのバイアスを有するかを識別するランキングとすることができる。

【００１９】

この方法は、前記多数のインテリジェンス・モデルをテストするためにテスト・データ・セットを受け取ることをさらに含む。テスト・データ・セットは、テスト分析結果と対になるテスト入力データの多数のグループを含む。テスト・データ・セットは、本質的に、多数の訓練される人工知能モデルの各々に入力するために使用される試行データである。特定のテスト・データ・セットでは、本質的に、人工知能モデルのうちの１つの正しいまたは所望の出力を提供するためにラベル付けされたグラウンド・トゥルースまたはデータであるテスト分析結果が存在する。

【００２０】

この方法は、前記テスト入力データの多数のグループを前記入力データ・セットとして入力することに応じて、前記多数の人工知能モデルの各々から緩和分析結果を受け取ることをさらに含む。緩和分析結果は、多数の人工知能モデルの試行の結果であると考えるこ

とができる。
【００２１】
この方法は、多数の訓練された人工知能モデルの各々に対する緩和分析結果と、テスト分析結果との間の比較を記述する、多数の訓練された人工知能モデルの各々に対する正確度スコアを計算することをさらに含む。
【００２２】
この方法は、１つまたは複数の選ばれた変数を試行分析結果と比較することによって、多数の訓練された人工知能モデルの各々に対する公平性評価メトリックを計算することをさらに含む。正確度スコアは、多数の訓練された人工知能モデルの各々がどれだけ正確であるかの尺度である。公平性評価メトリックは、多数の訓練された人工知能モデルの各々に対して、１つまたは複数の選ばれた変数にどれだけの望ましくないバイアスがあるかの尺度を提供する。
【００２３】
この方法は、次いで多数の訓練された人工知能モデルの各々に対して公平性評価メトリックと正確度スコアとを組み合わせることによって、多数の訓練された人工知能モデルの各々に対する公平性重み付きランキングを計算することを含む。そのため、正確度スコアを使用することによって多数の訓練された人工知能モデルをランキングする代わりに、正確度スコアと公平性評価メトリックの組み合わせが代わりに使用される。これは、モデルがどれだけ正確であるかの値だけでなく、どれだけ望ましくないバイアスが様々な人工知能モデルに存在するかの値も提供する。次いで、公平性重み付きランキングは、最良の人工知能モデルの自動選択に役立つことができ、またはユーザに表示されてもよく、ユーザは、公平性重み付きランキングに基づいてどのモデルを使用するかの選択を決定してもよい。
【００２４】
別の実施形態では、公平性評価メトリックは、前記１つまたは複数の選ばれた変数の１つまたは複数の選ばれた値と、試行分析結果との間の相関を記述する。例えば、公平性評価メトリックは、１つまたは複数の選ばれた変数の特定の値が差別されているかどうかを調べるために計算することができる。前記の例を使用して、特定のジェンダを選ぶことができ、この特定のジェンダが、訓練された人工知能モデルにバイアスをもたらすかどうかを調べることができる。これは、公平性評価メトリックが、訓練された人工知能モデルにおける特定のバイアスをチェックするために使用され得るので有益である。
【００２５】
別の実施形態では、多数の訓練された人工知能モデルは異なるタイプである。例えば、多数の訓練された人工知能モデルは、異なるニューラル・ネットワーク・トポロジを使用することができる。他の例では、異なるタイプは、人工知能の完全に異なる実施態様でさえあり得る。１つの例は、あるモデルがニューラル・ネットワークであり、他のモデルがベイズ決定モデルである場合であろう。この実施形態は、最良の人工知能トポロジまたはモデル・タイプあるいはその両方を選択することを可能にすることができるので有益であり得る。
【００２６】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つはニューラル・ネットワークである。
【００２７】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、分類器ニューラル・ネットワークである。
【００２８】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、畳み込みニューラル・ネットワークである。
【００２９】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、ベイジアン・ニ

ューラル・ネットワークである。
【００３０】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、ベイジアン・ネットワークである。
【００３１】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、ベイズ・ネットワークである。
【００３２】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、単純ベイズ分類器である。
【００３３】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、信念ネットワークである。
【００３４】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、決定ネットワークである。
【００３５】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、決定木である。
【００３６】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、サポート・ベクトル機械である。
【００３７】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、回帰分析である。
【００３８】
別の実施形態では、多数の訓練された人工知能モデルのうちの１つは、遺伝的アルゴリズムである。
【００３９】
別の実施形態では、公平性重み付きランキングは、公平性評価メトリックと正確度スコアの最小２乗組み合わせを含む。
【００４０】
別の実施形態では、公平性重み付きランキングは、公平性評価メトリックと正確度スコアの重み付き最小２乗組み合わせを含む。例えば、公平性評価メトリックを２乗し、次いで、第１の係数を乗算することができ、次いで、正確度スコアを２乗し、第２の係数を乗算し、次いで、２つのものを加算する。
【００４１】
別の実施形態では、公平性重み付きランキングは、評価メトリックと正確度スコアの線形結合を含む。
【００４２】
別の実施形態では、公平性重み付きランキングは、公平性評価メトリックと正確度スコアの重み付き組み合わせを含む。
【００４３】
別の実施形態では、公平性重み付きランキングは、公平性評価メトリックと正確度スコアの多項式組み合わせを含む。例えば、多項式は、様々な係数を用いて選ぶことができ、次いで、公平性評価メトリックおよび正確度スコアは、各々、異なる組み合わせで多項式に入れることができる。
【００４４】
別の実施形態では、組み合わせメトリックは、正確度スコアにスケーリング係数を乗算し、次いで、それを所定の冪乗したものである。スケーリング係数は、公平性評価メトリックの関数である。この実施形態は、公平性と正確度の良好な組み合わせ尺度を提供する

ことが示されているので有益であり得る。

【００４５】

別の実施形態では、スケーリング係数は、公平性評価メトリックの逆数である。

【００４６】

別の実施形態では、公平性スコア・メトリックは、前記１つまたは複数の選ばれた変数の１つまたは複数の選ばれた値と、試行分析結果との間の相関を記述する。公平性スコア・メトリックは、訓練中に人工知能モデルを評価するために使用される。この実施形態では、１つまたは複数の選ばれた値のうちの特定の値を選択することができ、これが差別されているかどうかまたは望ましくないバイアスを有しているかどうかを評価することができる。例えば、特定のジェンダに対する差別を避けるようにモデルを訓練することができる。

【００４７】

別の実施形態では、組み合わせメトリックは、公平性スコア・メトリックとテスト・メトリックの最小２乗組み合わせを含む。組み合わせメトリックは、公平性スコア・メトリックとテスト・メトリックの重み付き最小２乗組み合わせを含む。

【００４８】

別の実施形態では、組み合わせメトリックは、公平性スコア・メトリックとテスト・メトリックの線形組み合わせを含む。

【００４９】

別の実施形態では、組み合わせメトリックは、公平性スコア・メトリックとテスト・メトリックの重み付き組み合わせを含む。

【００５０】

別の実施形態では、組み合わせメトリックは、公平性スコア・メトリックとテスト・メトリックの多項式組み合わせを含む。

【００５１】

別の実施形態では、組み合わせメトリックは、公平性スコア・メトリックに対する制約を含む。例えば、制約は、公平性スコア・メトリックがどれだけ大きくなることが許容されるかに対して限定される可能性がある。これは、特定の変数に対してどれだけのバイアスがあるかに対して限定を有する訓練された人工知能モデルを提供することができる。

【００５２】

別の実施形態では、組み合わせメトリックは、テスト・メトリックに対する制約を含む。これは、例えば、訓練に許容できる最小正確度が存在するように訓練を限定するために使用することができるので、有用であり得る。これは、公平であるだけでなく正確でもあるモデルを構築するのに役立つことができる。

【００５３】

別の実施形態では、組み合わせメトリックは、公平性スコア・メトリックの最大許容値を含む。

【００５４】

別の実施形態では、組み合わせメトリックは、テスト・メトリックの最大許容値を含む。

【００５５】

別の実施形態では、人工知能モデルは、ニューラル・ネットワークである。

【００５６】

別の実施形態では、人工知能モデルは、分類器ニューラル・ネットワークである。

【００５７】

別の実施形態では、人工知能モデルは、畳み込みニューラル・ネットワークである。

【００５８】

別の実施形態では、人工知能モデルは、ベイジアン・ニューラル・ネットワークである。

【００５９】

別の実施形態では、人工知能モデルは、ベイジアン・ネットワークである。
【００６０】
別の実施形態では、人工知能モデルは、ベイズ・ネットワークである。
【００６１】
別の実施形態では、人工知能モデルは、単純ベイズ分類器である。
【００６２】
別の実施形態では、人工知能モデルは、信念ネットワークである。
【００６３】
別の実施形態では、人工知能モデルは、決定ネットワークである。
【００６４】
別の実施形態では、人工知能モデルは、決定木である。
【００６５】
別の実施形態では、人工知能モデルは、サポート・ベクトル機械である。
【００６６】
別の実施形態では、人工知能モデルは、回帰分析である。
【００６７】
別の実施形態では、人工知能モデルは、遺伝的アルゴリズムである。
【００６８】
別の実施形態では、人工知能モデルは、畳み込みニューラル・ネットワークである。訓練アルゴリズムは、深層学習アルゴリズムである。例えば、訓練アルゴリズムは、組み合わせメトリックを損失関数として使用する逆伝播アルゴリズムとすることができる。
【００６９】
本発明の実施形態は、コンピュータ・システム、クライアント、またはサーバとも呼ばれることがあるコンピューティング・デバイスを使用して実施することができる。次に図１を参照すると、コンピュータ・システムの一例の概略図が示される。コンピュータ・システム１０は、適切なコンピュータ・システムの単なる１つの例であり、本明細書に記載の本発明の実施形態の使用または機能の範囲に関していかなる限定も示唆するように意図されていない。それにもかかわらず、コンピュータ・システム１０は、上述に記載の機能のいずれかを実施または実行あるいはその両方を行うことができる。
【００７０】
コンピュータ・システム１０内には、非常に多くの他の汎用または専用コンピューティング・システム環境または構成により動作可能なコンピュータ・システム／サーバ１２がある。コンピュータ・システム／サーバ１２とともに使用するのに好適であり得るよく知られているコンピューティング・システム、環境、または構成、あるいはその組み合わせの例には、限定はしないが、パーソナル・コンピュータ・システム、サーバ・コンピュータ・システム、シン・クライアント、シック・クライアント、ハンドヘルドまたはラップトップ・デバイス、マルチプロセッサ・システム、マイクロプロセッサ・ベース・システム、セット・トップ・ボックス、プログラマブル家庭用電化製品、ネットワークＰＣ、ミニコンピュータ・システム、メインフレーム・コンピュータ・システム、および上述のシステムまたはデバイスのいずれかを含む分散コンピューティング環境、などが含まれる。
【００７１】
コンピュータ・システム／サーバ１２は、コンピュータ・システムによって実行されるプログラム・モジュールなどのコンピュータ・システム実行可能命令の一般的な文脈で説明することができる。一般に、プログラム・モジュールは、特定のタスクを実行するか、または特定の抽象データ型を実施するルーチン、プログラム、オブジェクト、コンポーネント、論理、データ構造、などを含むことができる。コンピュータ・システム／サーバ１２は、通信ネットワークを介してリンクされるリモート処理デバイスによってタスクが実行される分散コンピューティング環境において実践され得る。分散コンピューティング環境では、プログラム・モジュールは、メモリ・ストレージ・デバイスを含むローカルおよびリモートの両方のコンピュータ・システム・ストレージ媒体に配置することができる。

【００７２】
図１に示されるように、コンピュータ・システム１０内のコンピュータ・システム／サーバ１２は、汎用コンピューティング・デバイスの形態で示される。コンピュータ・システム／サーバ１２の構成要素は、限定はしないが、１つまたは複数のプロセッサまたは処理ユニット１６と、システム・メモリ２８と、システム・メモリ２８を含む様々なシステム構成要素をプロセッサ１６に結合させるバス１８とを含むことができる。バス１８は、メモリ・バスまたはメモリ・コントローラ、周辺バス、アクセラレーテッド・グラフィック・ポート、および様々なバス・アーキテクチャのうちのいずれかを使用するプロセッサまたはローカル・バスを含むいくつかのタイプのバス構造のいずれかの１つまたは複数を表す。例として、限定ではなく、そのようなアーキテクチャには、産業標準アーキテクチャ（ＩＳＡ）バス、マイクロ・チャネル・アーキテクチャ（ＭＣＡ）バス、エンハンストＩＳＡ（ＥＩＳＡ）バス、ビデオ・エレクトロニクス標準協会（ＶＥＳＡ）ローカル・バス、および周辺構成要素相互接続（ＰＣＩ）バスが含まれる。
【００７３】
コンピュータ・システム／サーバ１２は、一般に、様々なコンピュータ・システム可読媒体を含む。そのような媒体は、コンピュータ・システム／サーバ１２によってアクセス可能な任意の利用可能な媒体とすることができ、揮発性および不揮発性媒体と取り外し可能および取り外し不可媒体の両方を含む。
【００７４】
システム・メモリ２８は、ランダム・アクセス・メモリ（ＲＡＭ）３０またはキャッシュ・メモリ３２あるいはその両方などの揮発性メモリの形態のコンピュータ・システム可読媒体を含むことができる。コンピュータ・システム／サーバ１２は、他の取り外し可能／取り外し不可、揮発性／不揮発性のコンピュータ・システム・ストレージ媒体をさらに含むことができる。単なる例として、ストレージ・システム３４は、取り外し不可、不揮発性磁気媒体（図示せず、一般に、「ハード・ドライブ」と呼ばれる）から読み出し、それに書き込むために設けることができる。図示されていないが、取り外し可能、不揮発性磁気ディスク（例えば、「フロッピー（Ｒ）・ディスク」）から読み出し、それに書き込むための磁気ディスク・ドライブと、ＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭ、または他の光学媒体などの取り外し可能、不揮発性光ディスクから読み出し、またはそれに書き込むための光ディスク・ドライブとが設けられてもよい。そのような場合には、各々は、１つまたは複数のデータ媒体インターフェースによってバス１８に接続され得る。以下でさらに図示および説明されるように、メモリ２８は、本発明の実施形態の機能を実行するように構成されたプログラム・モジュールのセット（例えば、少なくとも１つ）を有する少なくとも１つのプログラム製品を含むことができる。
【００７５】
プログラム・モジュール４２のセット（少なくとも１つ）を有するプログラム／ユーティリティ４０は、例として、限定ではなく、オペレーティング・システム、１つまたは複数のアプリケーション・プログラム、他のプログラム・モジュール、およびプログラム・データと同様にメモリ２８に格納することができる。オペレーティング・システム、１つまたは複数のアプリケーション・プログラム、他のプログラム・モジュール、およびプログラム・データの各々、またはそれらの何らかの組み合わせは、ネットワーキング環境の実施態様を含むことができる。プログラム・モジュール４２は、通常、本明細書に記載の本発明の実施形態の機能または技法あるいはその両方を実行する。
【００７６】
コンピュータ・システム／サーバ１２はまた、キーボード、ポインティング・デバイス、ディスプレイ２４、などの１つまたは複数の外部デバイス１４、ユーザがコンピュータ・システム／サーバ１２と対話することを可能にする１つまたは複数のデバイス、またはコンピュータ・システム／サーバ１２が１つまたは複数の他のコンピューティング・デバイスと通信することを可能にする任意のデバイス（例えば、ネットワーク・カード、モデム、など）あるいはその組み合わせと通信することができる。そのような通信は、入力／

出力（Ｉ／Ｏ）インターフェース２２を介して行うことができる。さらに、コンピュータ・システム／サーバ１２は、ネットワーク・アダプタ２０を介して、ローカル・エリア・ネットワーク（ＬＡＮ）、汎用ワイド・エリア・ネットワーク（ＷＡＮ）、またはパブリック・ネットワーク（例えば、インターネット）、あるいはその組み合わせなどの１つまたは複数のネットワークと通信することができる。図示のように、ネットワーク・アダプタ２０は、バス１８を介してコンピュータ・システム／サーバ１２の他の構成要素と通信する。図示されていないが、他のハードウェア構成要素またはソフトウエア・コンポーネントあるいはその両方を、コンピュータ・システム／サーバ１２とともに使用することができることを理解されたい。例は、限定はしないが、マイクロコード、デバイス・ドライバ、冗長処理ユニット、外部ディスク・ドライブ・アレイ、ＲＡＩＤシステム、テープ・ドライブ、およびデータ・アーカイブ・ストレージ・システム、などを含む。

【００７７】

図１に示されるコンピュータ・システム１０などのコンピュータ・システムは、人工知能モジュールの訓練などの本明細書に開示される動作を実行するために使用することができる。そのようなコンピュータ・システムは、人工知能モジュールを訓練するための訓練データ・セットなどの処理されるべきデータをローカル・インターフェースを通して受け取ることができるネットワーク接続性のないスタンド・アロン・コンピュータとすることができる。しかしながら、そのような動作は、通信ネットワークまたはコンピューティング・ネットワークあるいはその両方などのネットワークに接続されるコンピュータ・システムを使用して、同様に実行されてもよい。

【００７８】

図２は、例示的なコンピューティング環境を示し、コンピュータ・システム１０などのコンピュータ・システムは、例えばネットワーク・アダプタ２０を使用してネットワーク２００に接続される。限定はしないが、ネットワーク２００は、インターネット、ローカル・エリア・ネットワーク（ＬＡＮ）、モバイル通信ネットワークなどの無線ネットワーク、などのような通信ネットワークとすることができる。ネットワーク２００は、クラウド・コンピューティング・ネットワークなどのコンピューティング・ネットワークを含むことができる。コンピュータ・システム１０は、人工知能モデルを訓練するための訓練データ・セットなどの処理されるべきデータをネットワーク２００から受け取ることができ、または訓練データ・セットを使用して訓練された後の訓練された人工知能モジュールなどのコンピューティング結果を、ネットワーク２００を介してコンピュータ・システム１０に接続された別のコンピューティング・デバイスに提供することができ、あるいはその両方である。

【００７９】

コンピュータ・システム１０は、ネットワーク２００を介して受け取った要求に応じて、本明細書に記載の動作を、完全にまたは部分的に、実行することができる。特に、コンピュータ・システム１０は、ネットワーク２００を介してコンピュータ・システム１０に接続され得る１つまたは複数のさらなるコンピュータ・システムとともに、そのような動作を分散計算で実行することができる。その目的のために、コンピューティング・システム１０または任意のさらなる関連するコンピュータ・システムあるいはその両方は、ネットワーク２００を使用して、専用メモリまたは共有メモリなどのさらなるコンピューティング・リソースにアクセスすることができる。

【００８０】

図３は、コンピュータ・システム１０の理想化されたものを示す。コンピュータ１０の処理ユニット１６またはプロセッサ、ならびにネットワーク・アダプタ２０およびＩ／Ｏインターフェース２２が示される。メモリ２８は、処理ユニット１６がアクセスすることができる様々なタイプのメモリを表す。処理ユニットは、機械実行可能命令３００を含むものとして示されている。機械実行可能命令３００は、プログラム・モジュール４２のうちの１つに相当する。メモリ２８の様々な内容は、様々な場所、例えば、ＲＡＭ３０、キャッシュ３２、または永続メモリなどに格納され得る。メモリ２８は、さらに、調整可能

パラメータを有する人工知能モデル３０２を含むものとして示されている。

【００８１】

人工知能モデルは、入力データ・セットを受け取ることに応じて、分析結果を提供するように訓練することができる。メモリ２８は、さらに、人工知能モデル３０２を訓練するために使用される訓練データ・セット３０４を含むものとして示されている。訓練データ・セット３０４は、訓練入力データ３０６の多数のグループのうちのグループと、訓練入力データの各々にとって利用可能であり得る訓練分析結果３０８とに分割することができる。訓練入力データ３０６は、人工知能モデル３０２に入力され、試行分析結果３１０を提供することができる。これは、メモリ２８に格納されるものとして示されている。

【００８２】

メモリ２８は、さらに、正確度メトリック３１２を含むものとして示されている。正確度メトリック３１２は、試行分析結果３１０と訓練分析結果３０８との間で計算された。メモリ２８は、さらに、入力データ・セットの１つまたは複数の選ばれた変数を試行分析結果３１０と比較することによって計算された公平性スコア・メトリック３１４を含むものとして示されている。メモリ２８は、さらに、公平性スコア・メトリック３１４と正確度メトリック３１２を組み合わせることによって計算された組み合わせメトリック３１６を含むものとして示されている。次いで、組み合わせメトリック３１６は、人工知能モデル３０２の調整可能パラメータを調節するために、訓練アルゴリズム３１８とともに使用される。

【００８３】

図４は、図３のコンピュータ１０を動作させる方法を示す流れ図を示す。最初に、ステップ４００において、訓練データ・セット３０４が受け取られる。次に、ステップ４０２において、試行分析結果３１０が、入力訓練データ３０６の多数のグループを入力データ・セットとして人工知能モデル３０２に入力することに応じて、人工知能モデル３０２から受け取られる。次に、ステップ４０４において、正確度メトリック３１２が計算され、それは、試行分析結果３１０と訓練分析結果３０８との間の比較を記述する。次いで、ステップ４０６において、公平性スコア・メトリック３１４が、１つまたは複数の選ばれた変数を試行分析結果３１０と比較することによって計算される。次に、ステップ４０８において、組み合わせメトリック３１６が、公平性スコア・メトリック３１４および正確度メトリック３１２から計算される。最後に、ステップ４１０において、人工知能モデル３０２の調節可能なパラメータが、少なくとも組み合わせメトリック３１６を入力として受け取る訓練アルゴリズム３１８を使用して修正される。

【００８４】

図５は、コンピュータ１０のさらなる図を示す。図３において示されたコンピュータ１０の特徴は、図５に示される特徴と組み合わせることができる。

【００８５】

メモリ２８は、機械実行可能命令３００を含むものとして示されている。メモリは、さらに、多数の訓練人工知能モデル５００を含むものとして示されている。図３に示された人工知能モデル３０２は、多分、多数の訓練された人工知能モデル５００のうちの１つとすることができるであろう。メモリ２８は、さらに、テスト入力データ５０４とテスト分析結果５０６とを含むテスト・データ・セット５０２を含むものとして示されている。テスト・データ・セット５０２は、多数の訓練された人工知能モデル５００をテストおよび評価するために使用される。テスト入力データ５０４は、入力として使用され、様々な人工知能モデルの出力が、テスト分析結果と比較される。

【００８６】

メモリ２８は、さらに、緩和分析結果を含むものとして示されている。緩和分析結果５０８は、テスト入力データが様々な人工知能モデルに入力されるときに、様々な人工知能モデルによって返される結果である。メモリ２８は、さらに、正確度スコア５１０を含むものとして示されている。正確度スコア５１０は、緩和分析結果５０８がテスト分析結果５０６に対してどれだけ正確であるかを評価するスコアである。メモリ２８は、さらに、

１つまたは複数の選ばれた変数を緩和分析結果５０８と比較することによって、多数の訓練された人工知能モデル５００の各々に対して計算された公平性評価メトリック５１２を含むものとして示されている。メモリ２８は、さらに、公平性重み付きランキング５１４を含むものとして示されている。公平性重み付きランキング５１４は、正確度スコア５１０と公平性評価メトリック５１２の組み合わせである。

【００８７】

図６は、図５のコンピュータ・システム１０を動作させる方法を示す流れ図を示す。図６に示される流れ図は、図４に示された流れ図と組み合わされてもよい。例えば、様々な人工モデルの訓練が図４に示された方法を使用して実行された後、多数の訓練された人工知能モデルは、図６に示される方法を使用して比較することができる。

【００８８】

最初に、ステップ６００において、多数の訓練された人工知能モデル５００が受け取られる。次に、ステップ６０２において、テスト・データ・セット５０２が受け取られる。次に、ステップ６０４において、緩和分析結果５０８が、様々な訓練された人工知能モデル５００にテスト入力データ５０４を入力することによって受け取られる。次に、ステップ６０６において、正確度スコア５１０が、多数の訓練された人工知能モデル５００の各々に対して、特定の知能モデル５００の緩和分析結果５０８と、テスト分析結果５０６とを比較することによって計算される。次に、ステップ６０８において、公平性評価メトリック５１２が、多数の訓練された人工知能モデル５００の各々に対して、１つまたは複数の選ばれた変数を緩和分析結果５０８と比較することによって計算される。最後に、ステップ６１０において、公平性重み付きランキング５１４が、多数の訓練された人工知能モデル５００の各々に対して、公平性評価メトリック５１２と正確度スコア５１０とを組み合わせることによって計算される。

【００８９】

自動機械学習手法は、今日では非常に普及している。それは、手動のデータ科学者の作業を自動化し、モデル開発プロセスを高速化することを可能にする。残念なことに、最良のモデルを見つけるには、かなりの量の時間およびリソースを必要とする可能性がある。自動機械学習プロセスの目標は、最も正確なモデルを見つけることである。

【００９０】

モデルが公平であることを確認することは、今日では、関連する可能性がある別の側面である。モデル公平性を評価し、緩和を可能にするように構成された専用の監視システムまたはライブラリがある。

【００９１】

実施形態は、バイアスチェックおよび緩和手順を自動機械学習プロセスに注入することができる。その手順は、スコアラ概念に基づく。

【００９２】

例示のシステムは、多分、２つのモジュール、すなわち、検出モジュール（人工知能モジュールの調節可能なパラメータを修正するために組み合わせメトリックを計算するために使用される）と、緩和モジュール（多数の訓練された人工知能モジュールに公平性評価メトリックを提供するための）とに基づくことができるであろう。モジュールは、別々にまたは一緒に使用することができる。

【００９３】

１．検出モジュール

【００９４】

検出モジュールは、公平性計算スコアラ（公平性評価メトリック）によって正規スコアラ・リストを拡張することに基づくことができる。スコアラ関数（本明細書では正確度スコアと呼ぶ）は、機械学習モデル（人工知能モデル）を評価するために使用される。サンプル・スコアラは、正確度、Ｂｒｉｅｒスコア損失、平均精度、バランス正確度、ｆ１スコア、などを含む。自動ＭＬ（ａｕｔｏＭＬ）プロセスの各段階中に、選択されたスコアラが探索プロセスを最適化するために使用され、その結果、最良のスコアラ値をもつモデ

ルが見いだされる。スコアラは探索プロセスを最適化するために使用され、その結果、最良のスコアラ値をもつモデルが見いだされる。スコアラは、モデルの性能（正確度）を記述する機械学習スコアラである。これらは、本明細書では「ｍｌ＿ｓｃｏｒｅｒｓ」と呼ばれる。

【００９５】

このモジュールには、公平性メトリック・スコアラ（公平性スコア・メトリック）をプロセスに追加することによるスコアラの拡張リストが存在する。言い換えれば、新しいタイプのスコアラが、既存のＭＬアーキテクチャに注入されている。これは、本明細書では、「ｆａｉｒｎｅｓｓ＿ｓｃｏｒｅｒ」または公平性スコア・メトリックと呼ばれる。ｍｌ＿ｓｃｏｒｅｒが計算されるたびに、「ｆａｉｒｎｅｓｓ＿ｓｃｏｒｅｒ」が（スコアラ・リストに追加されてから）、同様に実行され得る。

【００９６】

その結果、新しいメトリクスが、ユーザに返されることが可能であり、正確度、精度、および再現度などの機械学習メトリクスの次に、公平性スコア・メトリックが計算される。公平性スコア・メトリックは、本明細書ではｄｉｓｐａｒａｔｅ＿ｉｍｐａｃｔと呼ばれ、「ｆａｉｒｎｅｓｓ＿ｍｅｔｒｉｃｓ」カテゴリの下で計算される。

【００９７】

ｄｉｓｐａｒａｔｅ＿ｉｍｐａｃｔを計算するために、データ・セット内のあり得る困難（ａｄｖｅｒｓｉｔｙ）またはバイアスに関する情報が提供されてもよい。あり得るバイアスまたは偏見に関するこの情報は、本明細書では「ｆａｉｒｎｅｓｓ＿ｉｎｆｏ」と呼ばれる。ｆａｉｒｎｅｓｓ＿ｉｎｆｏの例および説明が以下で説明される。この情報は、パラメータとしてａｕｔｏＭＬシステムに渡され、公平性スコア・メトリックは、その情報に基づいて検出モジュールの各段階で計算される。公平性情報をもつ擬似コードにおけるシステムの例示的な呼出しは、以下のように示される。

```
〉〉〉 automl= AutoMLSystem(scorer= ‘accuracy’,
learning_type= ‘classification’,
positive_label= “No Risk”,
fairness_info= fairness_info
)
〉〉〉 automl.fit(training_data,training_labels)
```

【００９８】

‘ａｃｃｕｒａｃｙ’は、使用される正確度メトリックのタイプを指す。「ｔｒａｉｎｉｎｇ＿ｄａｔａ」は訓練入力データに対応し、「ｔｒａｉｎｉｎｇ＿ｌａｂｅｌｓ」は訓練分析結果に対応する。以下の公平性情報の保護された属性は、１つまたは複数の選ばれた変数に対応する。「ｐｒｏｔｅｃｔｅｄ＿ａｔｔｒｉｂｕｔｅｓ」の「ｐｒｉｖｉｌｅｇｅｄ＿ｇｒｏｕｐｓ」は、１つまたは複数の選ばれた変数の１つまたは複数の選ばれた値に対応する。

【００９９】

公平性情報の例：

【０１００】

－　分類：　「ｐｒｉｖｉｌｅｇｅｄ＿ｇｒｏｕｐｓ」の分類の以下の例は、バイアスされることがある１つまたは複数の選ばれた変数の特定の値であり得る。

```
fairness_info= {
 “protected_attributes”: [
{“feature”: “Gender”, “privileged_groups”: [‘male’]},
{“feature”: “Age”, “privileged_groups”: [[0.0, 40.0]]},
 ],
“favorable_labels”: [“No Risk”]}
```

【０１０１】

－　回帰

```
fairness_info= {
“favorable_labels”: [[-100000.0, 100]],
“protected_attributes”:[
{“feature”: “B”, “privileged_groups”: [[0.0, 40.0]]},
]}
```

ここで、

－ protected_attribute（アイテムのディクショナリ） － 公平性が望ましい特徴名および特権グループのサブセット。

－ favorable_labels（アレイ） － 好ましい（すなわち、「肯定的である」）と考えられるラベル値。利用可能なタイプ：ストリング、数、数のアレイ

【０１０２】

例示的なメトリクス出力：

パイプライン０のスコア：異種の影響：０．８１、正確度および異種の影響：０．７１

パイプライン１のスコア：異種の影響：０．８４、正確度および異種の影響：０．７７

パイプライン２のスコア：異種の影響：０．６７、正確度および異種の影響：０．８２

パイプライン３のスコア：異種の影響：０．６６、正確度および異種の影響：０．８４

【０１０３】

上述において、「異種の影響」は「公平性スコア・メトリック」であり、「正確度および異種の影響」は「組み合わせメトリック」である。

【０１０４】

２．緩和モジュール

【０１０５】

緩和モジュールは、再び、スコアラ手法に基づく。ここで、いわゆる組み合わせスコアラが、もう一度導入される。いくつかの重みに基づいてＭＬ（正確度スコア）と公平性メトリック（公平性評価メトリック）の両方を組み合わせた組み合わせスコアラは、本明細書では公平性重み付きランキングまたは本明細書では「accuracy_and_disparate_impact_scorer」とも呼ばれる。次に、そのようなスコアラは、ランキング・スコアラとして設定され、最適化プロセスで使用される。これは、計算されたスコア値（公平性重み付きランキング）に従って最良のモデルを見つけることに関与するプロセスである。緩和モジュールでは、それは、組み合わされた値である。それは、ａｕｔｏＭＬシステムの各段階で計算されるが、加えて、モデル選択ステップ中のモデル・ランキングのために使用される（多数の訓練された人工知能モデルの各々に公平性重み付きランキングを提供することによって）。組み合わせスコアラのうちの１つは、公平性スコアラ（公平性評価メトリック）であり、検出モジュールの公平性スコア・メトリックと類似しており、それは、すべてのａｕｔｏＭＬシステム・ステップで計算することができ、提供されたｆａｉｒｎｅｓｓ_ｉｎｆｏも使用する。組み合わせスコアラの最終値は、異種の影響の比に依存する。

【０１０６】

公平性メトリック（公平性評価メトリック）がＮａＮ（ゼロ除算によって生じるものなどの非数（ｎｏｔ ａ ｎｕｍｂｅｒ））である場合、公平性情報は、データ・セット・サンプル（例えば、ｋ－分割交差検証からのサンプル）に適していないので、組み合わせメトリックからの第２のメトリック、例えば正確度が返される。

【０１０７】

異種の影響の比（公平性評価メトリック）が０．０に等しいとき、組み合わせメトリック（公平性重み付きランキング）の最終値は０．０である。

【０１０８】

そうでない場合、組み合わせメトリックは、以下の式を使用して、両方のメトリックの混合として計算される。

【０１０９】

正確度（正確度スコア）および異種の影響（公平性評価メトリック）＝正確度＊（スケ

ーリング係数）^（スケーリング硬度）
【０１１０】
ここで、
【０１１１】
スケーリング係数は、０．９に設定されたパラメータである異種の影響の閾値（この閾値を超える値は公平と考えられる）と、以下で説明するパラメータである対称影響値（ｓｙｍｍｅｔｒｉｃ　ｉｍｐａｃｔ　ｖａｌｕｅ）とに依存する。異種の影響が０と０．９の間にある場合、対称影響は、異種の影響に等しい。異種の影響が１．０より大きい場合、対称影響は、以下の式を使用して計算される。
【０１１２】
ｓｃａｌｉｎｇ＿ｆａｃｔｏｒ＝（対称影響）／（異種の影響の閾値）
【０１１３】
スケーリング硬度は、４．０に設定されたパラメータである。
【０１１４】
例示的な緩和モジュールにおいて利用可能な２つの組み合わせスコアラ（公平性重み付きランキングを計算するための）がある。
【０１１５】
－　回帰：　ｒ２＿ａｎｄ＿ｄｉｓｐａｒａｇｅ＿ｉｍｐａｃｔ
【０１１６】
－　分類：ａｃｃｕｒａｃｙ＿ａｎｄ＿ｄｉｓｐａｒａｔｅ＿ｉｍｐａｃｔ
【０１１７】
緩和モジュールの例示的な呼出し：

```
〉〉〉 automl = AutoMLSystem(scorer = ‘accuracy_and_disparate_impact’,
learning_type= ‘classification’
positive_label= “No Risk”,
fairness_info= fairness_info
)
〉〉〉 automl.fit(training_data,training_labels)
```

【０１１８】
公平性情報の例：
－　分類

```
Fairness_info ={
 “protected_attributes”:[
{“feature”: “GENDER”, “privileged_groups”; [‘F’]},
{“feature”: “BP”, “privileged_groups”: [“LOW”, “NORMAL”]}
 ],
“favorable_labels”: [“drugA”, “durgC”]}
```

【０１１９】
例示的なメトリクス出力：
パイプライン０のスコア：異種の影響：０．６０、正確度および異種の影響：０．６４
パイプライン１のスコア：異種の影響：０．６６、正確度および異種の影響：０．６８
パイプライン２のスコア：異種の影響：０．７１、正確度および異種の影響：０．７７
パイプライン３のスコア：異種の影響：０．７０、正確度および異種の影響：０．８１
【０１２０】
上述において、「異種の影響」は「公平性評価メトリック」であり、「正確度および異種の影響」は「公平性重み付きランキング」である。
【０１２１】
モデル・ランキングはまた、解釈を容易にするために、両方のメトリック（分離された）、すなわち、正確度のような機械学習メトリックと異種の影響のような公平性メトリックとを使用して行うことができる。それは、エンド・ユーザへの有用な提示と、選択され

たメトリックに基づいてランキングまたはソートあるいはその両方を行う能力とを可能にする。

【０１２２】

その選択はまた、いくつかの閾値に基づくフィルタリングに容易に拡張することができる。ユーザは、例えば、最良の公平性パイプラインを提供するが、精度が０．８以上であるという制約を設定する。

【０１２３】

本発明は、任意の可能な技術的詳細レベルの統合におけるシステム、方法、またはコンピュータ・プログラム製品、あるいはその組み合わせであり得る。コンピュータ・プログラム製品は、プロセッサに本発明の態様を実行させるためのコンピュータ可読プログラム命令を有する１つのコンピュータ可読ストレージ媒体（または複数の媒体）を含む。

【０１２４】

コンピュータ可読ストレージ媒体は、命令実行デバイスによる使用のための命令を保持および格納することができる有形のデバイスとすることができる。コンピュータ可読ストレージ媒体は、例えば、限定はしないが、電子ストレージ・デバイス、磁気ストレージ・デバイス、光学ストレージ・デバイス、電磁気ストレージ・デバイス、半導体ストレージ・デバイス、または前述のものの任意の適切な組み合わせとすることができる。コンピュータ可読ストレージ媒体のより具体的な例の非網羅的なリストは、以下のもの、すなわち、ポータブル・コンピュータ・ディスケット、ハード・ディスク、ランダム・アクセス・メモリ（ＲＡＭ）、読み取り専用メモリ（ＲＯＭ）、消去可能プログラマブル読み取り専用メモリ（ＥＰＲＯＭまたはフラッシュ・メモリ）、スタティック・ランダム・アクセス・メモリ（ＳＲＡＭ）、ポータブル・コンパクト・ディスク読み取り専用メモリ（ＣＤ－ＲＯＭ）、デジタル・バーサタイル・ディスク（ＤＶＤ）、メモリ・スティック、フロッピー（Ｒ）・ディスク、命令が記録されたパンチカードまたは溝内の隆起構造などの機械的にコード化されたデバイス、および前述のものの任意の適切な組み合わせを含む。本明細書で使用されるコンピュータ可読ストレージ媒体は、電波もしくは他の自由に伝播する電磁波、導波路もしくは他の伝送媒体を通って伝搬する電磁波（例えば、光ファイバ・ケーブルを通過する光パルス）、またはワイヤを通して伝送される電気信号などのそれ自体一過性信号であると解釈されるべきではない。

【０１２５】

本明細書に記載のコンピュータ可読プログラム命令は、コンピュータ可読ストレージ媒体からそれぞれのコンピューティング／処理デバイスに、あるいはネットワーク、例えば、インターネット、ローカル・エリア・ネットワーク、ワイド・エリア・ネットワーク、または無線ネットワーク、あるいはその組み合わせを介して外部コンピュータまたは外部ストレージ・デバイスにダウンロードされ得る。ネットワークは、銅伝送ケーブル、光伝送ファイバ、無線伝送、ルータ、ファイアウォール、スイッチ、ゲートウェイ・コンピュータ、またはエッジ・サーバ、あるいはその組み合わせを含むことができる。各コンピューティング／処理デバイスのネットワーク・アダプタ・カードまたはネットワーク・インターフェースは、コンピュータ可読プログラム命令をネットワークから受け取り、そのコンピュータ可読プログラム命令をそれぞれのコンピューティング／処理デバイス内のコンピュータ可読ストレージ媒体に格納するために転送する。

【０１２６】

本発明の動作を実行するためのコンピュータ可読プログラム命令は、アセンブラ命令、命令セット・アーキテクチャ（ＩＳＡ）命令、機械命令、機械依存命令、マイクロコード、ファームウェア命令、状態設定データ、集積回路のための構成データ、あるいはＳｍａｌｌｔａｌｋ（Ｒ）、Ｃ＋＋などのようなオブジェクト指向プログラミング言語および「Ｃ」プログラミング言語または同様のプログラミング言語などの手続き型プログラミング言語を含む１つまたは複数のプログラミング言語の任意の組み合わせで書かれたソース・コードまたはオブジェク・トコードのいずれかとすることができる。コンピュータ可読プログラム命令は、全面的にユーザのコンピュータで、部分的にユーザのコンピュータで、

スタンドアロン・ソフトウェア・パッケージとして、部分的にユーザのコンピュータでおよび部分的にリモート・コンピュータで、または全面的にリモート・コンピュータもしくはサーバで実行することができる。後者のシナリオでは、リモート・コンピュータは、ローカル・エリア・ネットワーク（ＬＡＮ）もしくはワイド・エリア・ネットワーク（ＷＡＮ）を含む任意のタイプのネットワークを通してユーザのコンピュータに接続されてもよく、または接続が外部コンピュータに対して行われてもよい（例えば、インターネット・サービス・プロバイダを使用してインターネットを通して）。いくつかの実施形態では、例えば、プログラマブル論理回路、フィールド・プログラマブル・ゲート・アレイ（ＦＰＧＡ）、またはプログラマブル論理アレイ（ＰＬＡ）を含む電子回路は、本発明の態様を実行するために、コンピュータ可読プログラム命令の状態情報を利用して電子回路を個人専用にすることによってコンピュータ可読プログラム命令を実行することができる。

【０１２７】

本発明の態様は、本発明の実施形態による方法、装置（システム）、およびコンピュータ・プログラム製品の流れ図またはブロック図あるいはその両方を参照して本明細書に記載される。流れ図またはブロック図あるいはその両方の各ブロック、および流れ図またはブロック図あるいはその両方におけるブロックの組み合わせは、コンピュータ可読プログラム命令によって実現され得ることが理解されよう。

【０１２８】

これらのコンピュータ可読プログラム命令は、汎用コンピュータ、専用コンピュータ、または他のプログラマブル・データ処理装置のプロセッサに提供されて、コンピュータまたは他のプログラマブル・データ処理装置のプロセッサを介して実行される命令が流れ図またはブロック図あるいはその両方の１つまたは複数のブロックにおいて指定された機能／動作を実施するための手段を作り出すような機械を生成することができる。これらのコンピュータ可読プログラム命令はまた、命令が格納されたコンピュータ可読ストレージ媒体が流れ図またはブロック図あるいはその両方の１つまたは複数のブロックにおいて指定された機能／動作の態様を実施する命令を含む製品を構成するように、コンピュータ、プログラマブル・データ処理装置、または他のデバイス、あるいはその組み合わせに、特定のやり方で機能するように指示することができるコンピュータ可読ストレージ媒体に格納されてもよい。

【０１２９】

コンピュータ可読プログラム命令はまた、コンピュータ、他のプログラマブル・データ処理装置、または他のデバイスにロードされて、一連の動作ステップをコンピュータ、他のプログラマブル装置、または他のデバイスで実行させて、コンピュータ実施プロセスを作り出し、その結果、コンピュータ、他のプログラマブル装置、または他のデバイスで実行される命令が流れ図またはブロック図あるいはその両方の１つまたは複数のブロックにおいて指定された機能／動作を実施することができる。

【０１３０】

図における流れ図およびブロック図は、本発明の様々な実施形態によるシステム、方法、およびコンピュータ・プログラム製品の可能な実施態様のアーキテクチャ、機能、および動作を示す。これに関しては、流れ図またはブロック図の各ブロックは、指定された論理機能を実施するための１つまたは複数の実行可能命令を含む命令のモジュール、セグメント、または一部を表すことができる。いくつかの代替実施態様では、ブロックに記された機能は、図に記された順序から外れて行われてもよい。例えば、連続して示された２つのブロックは、実際には、実質的に同時に実行されてもよく、またはブロックは、時には、関連する機能に応じて逆の順序で実行されてもよい。ブロック図または流れ図あるいはその両方の各ブロック、およびブロック図または流れ図あるいはその両方のブロックの組み合わせは、指定された機能または動作を実行するかあるいは専用ハードウェア命令とコンピュータ命令の組み合わせを実行する専用ハードウェア・ベース・システムで実施され得ることにも留意されたい。

【０１３１】

様々な例は、多分、以下の番号付けされた条項における以下の特徴のうちの１つまたは複数によって記載され得るであろう。

【０１３２】

条項１。

人工知能モデルを訓練する方法であって、人工知能モデルが、調節可能なパラメータを有し、前記人工知能モデルが、入力データ・セットを受け取ることに応じて、分析結果を提供するように訓練され、前記入力データ・セットが、１つまたは複数の選ばれた変数を含み、前記方法が、

前記人工知能モデルを訓練するための訓練データ・セットを受け取ることであり、前記訓練データ・セットが、訓練分析結果と対になる訓練入力データの多数のグループを含む、受け取ることと、

前記訓練入力データの多数のグループを前記入力データ・セットとして前記人工知能モデルに入力することに応じて、前記人工知能モデルから試行分析結果を受け取ることと、

前記試行分析結果と前記訓練分析結果との間の比較を記述する正確度メトリックを計算することと、

前記１つまたは複数の選ばれた値を前記試行分析結果と比較することによって公平性スコア・メトリックを計算することと、

前記公平性スコア・メトリックおよび前記正確度メトリックから組み合わせメトリックを計算することと、

少なくとも前記組み合わせメトリックを入力として受け取る訓練アルゴリズムを使用して、人工知能モデルの調節可能なパラメータを修正することと

を含む、方法。

【０１３３】

条項２。

前記方法は、

前記多数の訓練された人工知能モデルを受け取ることであり、前記多数の訓練された人工知能モデルが前記人工知能モデルからなる、受け取ることと、

前記多数の人工知能モデルをテストするためのテスト・データ・セットを受け取ることであり、前記テスト・データ・セットが、テスト分析結果と対になるテスト入力データの多数のグループを含む、受け取ることと、

前記テスト入力データの多数のグループを前記入力データ・セットとして入力することに応じて、前記多数の人工知能モデルの各々から緩和分析結果を受け取ることと、

前記多数の訓練された人工知能モデルの各々に対する前記緩和分析結果と、前記テスト分析結果との間の比較を記述する、前記多数の訓練された人工知能モデルの各々に対する正確度スコアを計算することと、

前記１つまたは複数の選ばれた値を前記試行分析結果と比較することによって、前記多数の訓練された人工知能モデルの各々に対する公平性評価メトリックを計算することと、

多数の訓練された人工知能モデルの各々に対して前記公平性評価メトリックと正確度スコアとを組み合わせることによって、前記多数の訓練された人工知能モデルの各々に対する前記公平性重み付きランキングを計算することと

によって、多数の訓練された人工知能モデルの各々に対する公平性重み付きランキングを提供することをさらに含む、条項１に記載の方法。

【０１３４】

条項３。

前記公平性評価メトリックが、前記１つまたは複数の選ばれた変数の１つまたは複数の選ばれた値と、前記試行分析結果との間の相関を記述する、条項２に記載の方法。

【０１３５】

条項４。

前記多数の訓練された人工知能モデルが、異なるタイプのものである、条項２または３に記載の方法。

【０１３６】
条項５。
前記多数の訓練された人工知能モデルの各々が、独立して、以下のもの、すなわち、ニューラル・ネットワーク、分類器ニューラル・ネットワーク、畳み込みニューラル・ネットワーク、ベイジアン・ニューラル・ネットワーク、ベイジアン・ネットワーク、ベイズ・ネットワーク、単純ベイズ分類器、信念ネットワーク、または決定ネットワーク、決定木、サポート・ベクトル機械、回帰分析、および遺伝的アルゴリズムのうちの任意の１つである、条項２、３、または４に記載の方法。
【０１３７】
条項６。
前記公平性重み付きランキングが、以下のもの、すなわち、前記公平性評価メトリックと前記正確度スコアの最小２乗組み合わせ、前記公平性評価メトリックと前記正確度スコアの重み付き最小２乗組み合わせ、前記公平性評価メトリックと前記正確度スコアの線形組み合わせ、前記公平性評価メトリックと前記正確度スコアの重み付き組み合わせ、および前記公平性評価メトリックと前記正確度スコアの多項式組み合わせのうちの任意の１つを含む、条項１ないし５のいずれか一項に記載の方法。
【０１３８】
条項７。
前記組み合わせメトリックが、前記正確度スコアにスケーリング係数を乗算し、それを所定の冪乗したものであり、前記スケーリング係数が、前記公平性評価メトリックの関数である、条項２ないし５のいずれか一項に記載の方法。
【０１３９】
条項８。
前記スケーリング係数が、前記公平性評価メトリックの逆数である、条項７に記載の方法。
【０１４０】
条項９。
前記公平性スコア・メトリックが、前記１つまたは複数の選ばれた変数の１つまたは複数の選ばれた値と、前記試行分析結果との間の相関を記述する、条項１ないし８のいずれか一項に記載の方法。
【０１４１】
条項１０。
前記組み合わせメトリックが、以下のもの、すなわち、前記公平性スコア・メトリックと前記テスト・メトリックの最小２乗組み合わせ、前記公平性スコア・メトリックと前記テスト・メトリックの重み付き最小２乗組み合わせ、前記公平性スコア・メトリックと前記テスト・メトリックの線形組み合わせ、前記公平性スコア・メトリックと前記テスト・メトリックの重み付き組み合わせ、および前記公平性スコア・メトリックと前記テスト・メトリックの多項式組み合わせのうちの任意の１つを含む、条項１ないし９のいずれか一項に記載の方法。
【０１４２】
条項１１。
前記組み合わせメトリックが、以下のもの、すなわち、前記公平性スコア・メトリックに対する制約、前記テスト・メトリックに対する制約、前記公平性スコア・メトリックに対する最大許容値、および前記テスト・メトリックに対する最大許容値のうちの任意の１つを含む、条項９または１０に記載の方法。
【０１４３】
条項１２。
前記人工知能モデルが、以下のもの、すなわち、ニューラル・ネットワーク、分類器ニューラル・ネットワーク、畳み込みニューラル・ネットワーク、ベイジアン・ニューラル・ネットワーク、ベイジアン・ネットワーク、ベイズ・ネットワーク、単純ベイズ分類器

、信念ネットワーク、または決定ネットワーク、決定木、サポート・ベクトル機械、回帰分析、および遺伝的アルゴリズムのうちの任意の１つである、条項１ないし１１のいずれか一項に記載の方法。

【０１４４】

条項１３。

前記人工知能モデルが畳み込みニューラル・ネットワークであり、前記訓練アルゴリズムが深層学習アルゴリズムである、条項１ないし１２のいずれか一項に記載の方法。

【０１４５】

条項１４。

コンピュータ可読プログラム・コードが具現化されたコンピュータ可読ストレージ媒体を含むコンピュータ・プログラム製品であって、前記コンピュータ可読のプログラム・コードが、条項１ないし１３のいずれか一項に記載の方法を実施するように構成される、コンピュータ・プログラム製品。

【０１４６】

条項１５。

コンピュータ・システムを制御するように構成されたプロセッサと、

機械実行命令を格納するメモリであり、前記命令の実行により、前記プロセッサが、

人工知能モデルを訓練するための訓練データ・セットを受け取ることであり、人工知能モデルが、調節可能なパラメータを有し、前記人工知能モデルが、入力データ・セットを受け取ることに応じて、分析結果を提供するように訓練され、前記入力データ・セットが、１つまたは複数の選ばれた変数を含み、前記訓練データ・セットが、訓練分析結果と対になる訓練入力データの多数のグループを含む、受け取ることと、

前記訓練入力データの多数のグループを前記入力データ・セットとして前記人工知能モデルに入力することに応じて、前記人工知能モデルから試行分析結果を受け取ることと、

前記試行分析結果と前記訓練分析結果との間の比較を記述する正確度メトリックを計算することと、

前記１つまたは複数の選ばれた値を前記試行分析結果と比較することによって計算される公平性スコア・メトリックを計算することと、

前記公平性スコア・メトリックおよび前記正確度メトリックから組み合わせメトリックを計算することと、

少なくとも前記組み合わせメトリックを入力として受け取る訓練アルゴリズムを使用して、人工知能モデルの調節可能なパラメータを修正することと

を行う、メモリと

を含むコンピュータ・システム。

【０１４７】

条項１６。

命令の実行により、さらに、前記プロセッサが、

前記多数の訓練された人工知能モデルを受け取ることであり、前記多数の訓練された人工知能モデルが前記人工知能モデルからなる、受け取ることと、

前記多数の人工知能モデルをテストするためのテスト・データ・セットを受け取ることであり、前記テスト・データ・セットが、テスト分析結果と対になるテスト入力データの多数のグループを含む、受け取ることと、

前記テスト入力データの多数のグループを前記入力データ・セットとして入力することに応じて、前記多数の人工知能モデルの各々から緩和分析結果を受け取ることと、

前記多数の訓練された人工知能モデルの各々に対する前記緩和分析結果と、前記テスト分析結果との間の比較を記述する、前記多数の訓練された人工知能モデルの各々に対する正確度スコアを計算することと、

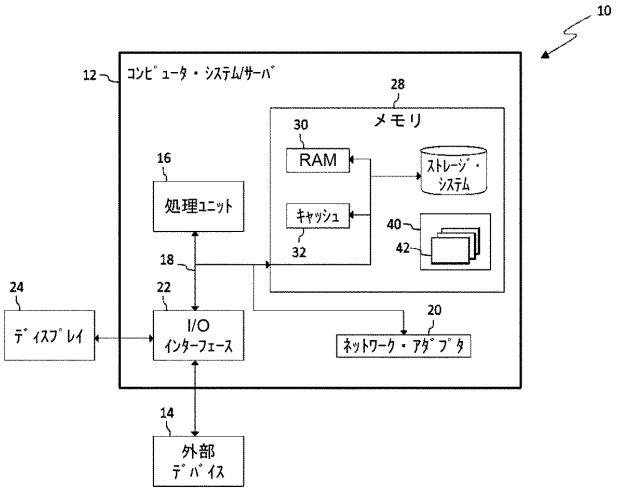
前記１つまたは複数の選ばれた値を前記試行分析結果と比較することによって、前記多数の訓練された人工知能モデルの各々に対する公平性評価メトリックを計算することと、

多数の訓練された人工知能モデルの各々に対して前記公平性評価メトリックと正確度スコアとを組み合わせることによって、前記多数の訓練された人工知能モデルの各々に対する前記公平性重み付きランキングを計算することと
を行う、条項１５に記載のコンピュータ・システム。
【０１４８】
条項１７。
人工知能モデルが、以下のもの、すなわち、ニューラル・ネットワーク、分類器ニューラル・ネットワーク、畳み込みニューラル・ネットワーク、ベイジアン・ニューラル・ネットワーク、ベイジアン・ネットワーク、ベイズ・ネットワーク、単純ベイズ分類器、信念ネットワーク、または決定ネットワーク、決定木、サポート・ベクトル機械、回帰分析、および遺伝的アルゴリズムのうちの任意の１つである、条項１５ないし１６のいずれか一項に記載のコンピュータ・システム。
【０１４９】
条項１８。
前記人工知能モデルが畳み込みニューラル・ネットワークであり、前記訓練アルゴリズムが深層学習アルゴリズムである、条項１５ないし１７のいずれか一項に記載のコンピュータ・システム。
【０１５０】
条項１９。
コンピュータ・プログラム製品であって、前記コンピュータ・プログラム製品が、条項１ないし１２のいずれか一項に記載の方法に従って訓練された人工知能モデルを格納したコンピュータ可読ストレージ媒体を含む、コンピュータ・プログラム製品。
【０１５１】
条項２０。
データ処理システムで実行されるアプリケーション・プログラムがアクセスするためのデータを格納するためのメモリであって、条項１ないし１２のいずれか一項に記載の方法に従って訓練される人工知能モデルを含む、メモリ。
【０１５２】
条項２１。
多数の訓練された人工知能モデルの各々に公平性重み付きランキングを提供する方法であって、この方法が、
前記多数の訓練された人工知能モデルを受け取ることと、
前記多数の人工知能モデルをテストするためのテスト・データ・セットを受け取ることであり、前記テスト・データ・セットが、テスト分析結果と対になるテスト入力データの多数のグループを含む、受け取ることと、
前記テスト入力データの多数のグループを前記入力データ・セットとして入力することに応じて、前記多数の人工知能モデルの各々から緩和分析結果を受け取ることと、
前記多数の訓練された人工知能モデルの各々に対する前記緩和分析結果と、前記テスト分析結果との間の比較を記述する、前記多数の訓練された人工知能モデルの各々に対する正確度スコアを計算することと、
前記１つまたは複数の選ばれた値を前記試行分析結果と比較することによって、前記多数の訓練された人工知能モデルの各々に対する公平性評価メトリックを計算することと、
多数の訓練された人工知能モデルの各々に対して前記公平性評価メトリックと正確度スコアとを組み合わせることによって、前記多数の訓練された人工知能モデルの各々に対する前記公平性重み付きランキングを計算することと
を含む、方法。
【０１５３】
本発明の様々な実施形態の説明は、例証の目的のために提示されたが、網羅的であること、または開示された実施形態に限定されることを意図するものではない。説明された実施形態の範囲および思想から逸脱することなく、多くの変更および変形が当業者には明ら

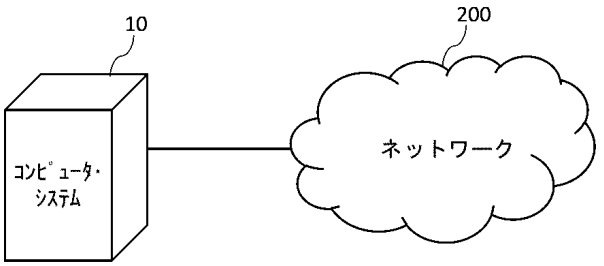
かであろう。本明細書で使用される用語は、実施形態の原理、実際の適用、もしくは市場で見いだされる技術に対する技術的改善を最もよく説明するために、または他の当業者が本明細書に開示される実施形態を理解できるようにするために選ばれた。

【図面】

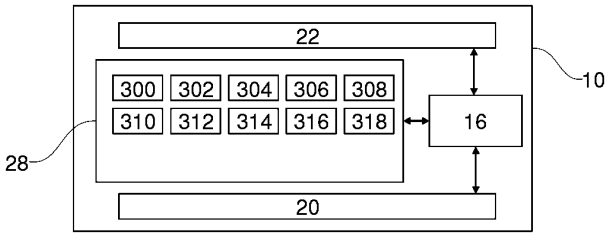
【図１】



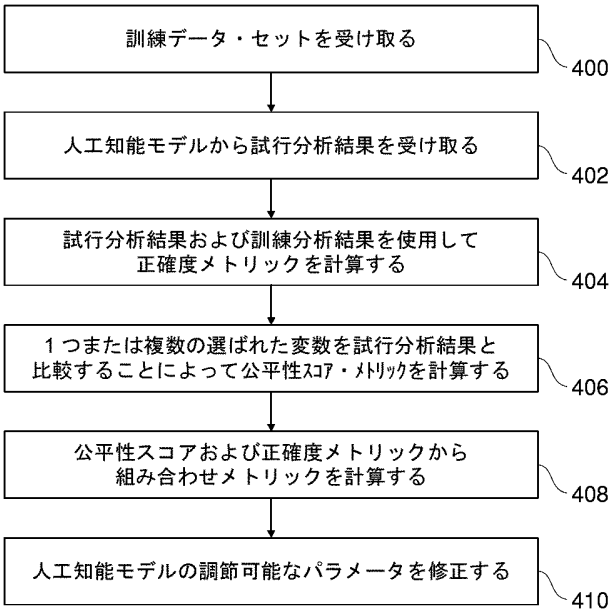
【図２】



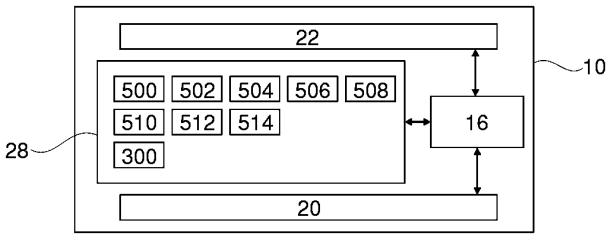
【図３】



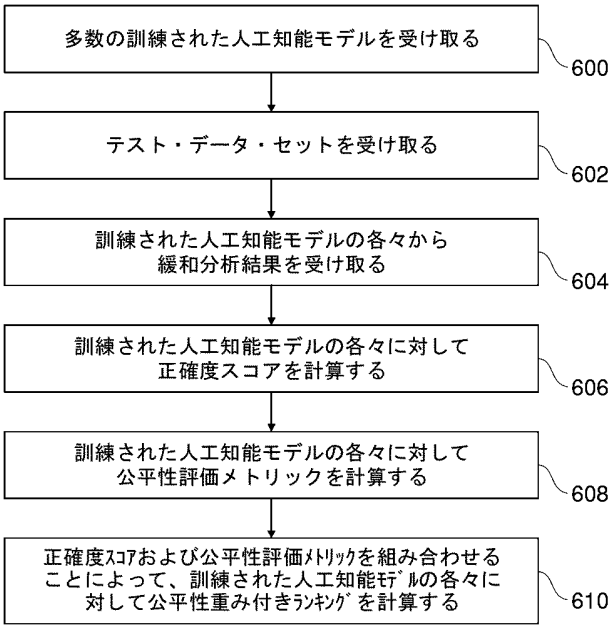
【図４】



【図５】


22
500
502
504
506
508
510
512
514
300
16
20
10
28

【図６】


多数の訓練された人工知能モデルを受け取る
600
テスト・データ・セットを受け取る
602
訓練された人工知能モデルの各々から
緩和分析結果を受け取る
604
訓練された人工知能モデルの各々に対して
正確度スコアを計算する
606
訓練された人工知能モデルの各々に対して
公平性評価メトリックを計算する
608
正確度ｽｺｱおよび公平性評価ﾒﾄﾘｯｸを組み合わせる
ことによって、訓練された人工知能ﾓﾃﾞﾙの各々に
対して公平性重み付きﾗﾝｷﾝｸﾞを計算する
610

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2022/055104

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08(2006.01)i; G06N 20/00(2019.01)i; G06N 7/00(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N3/-,G06N20/-,G06N7/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, CNABS, DWPI, CNKI, WPABS, ENTXT, ENTXTC:AI,artificial w intelligence, model, deep 3d learning, neural 3d network, class, classifier, attribute, race, ethnicity, age, sex, dateset, data w set,input, output, variable, adjustable w parameter+, training, fair, accuracy, fairness, metric+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2020302524 A1 (ZESTFINANCE INC.) 24 September 2020 (2020-09-24) description, paragraphs [0042], [0048] to [0136] and figures 1A to 7	1-21
A	US 2020320428 A1 (IBM) 08 October 2020 (2020-10-08) the whole document	1-21
A	CN 110782004 A (CHAOCANSHU TECHNOLOGY SHENZHEN CO., LTD.) 11 February 2020 (2020-02-11) the whole document	1-21
A	CN 112541579 A (BEIJING BEIMING SHUKE INFORMATION TECHNOLOGY CO., LTD.) 23 March 2021 (2021-03-23) the whole document	1-21
A	US 2020372406 A1 (ORACLE INTERNATIONAL CORP.) 26 November 2020 (2020-11-26) the whole document	1-21
A	US 2021158204 A1 (IBM) 27 May 2021 (2021-05-27) the whole document	1-21
A	US 2020342307 A1 (IBM) 29 October 2020 (2020-10-29) the whole document	1-21

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 July 2022	**29 August 2022**
Name and mailing address of the ISA/CN **National Intellectual Property Administration, PRC** **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing** **100088, China**	Authorized officer **HUANG,Bin**
Facsimile No. **(86-10)62019451**	Telephone No. **86-(10)-53962532**

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2022/055104

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2020226489 A1 (ADOBE INC.) 16 July 2020 (2020-07-16) the whole document	1-21

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/IB2022/055104

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 2020302524 A1	24 September 2020	WO 2020191057 A1	24 September 2020
		US 2021133870 A1	06 May 2021
		CA 3134043 A1	24 September 2020
		EP 3942384 A1	26 January 2022
		JP 2022525702 A	18 May 2022
US 2020320428 A1	08 October 2020	WO 2020208444 A1	15 October 2020
		CN 113692594 A	23 November 2021
		JP 2022527536 A	02 June 2022
		DE 112020000537 T5	21 October 2021
		GB 2597406 A	26 January 2022
CN 110782004 A	11 February 2020	None	
CN 112541579 A	23 March 2021	None	
US 2020372406 A1	26 November 2020	None	
US 2021158204 A1	27 May 2021	None	
US 2020342307 A1	29 October 2020	None	
US 2020226489 A1	16 July 2020	None	

Form PCT/ISA/210 (patent family annex) (January 2015)

フロントページの続き

MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IQ,IR,IS,IT,JM,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

(74)代理人 100120710
弁理士 片岡 忠彦
(72)発明者 クミエロウスキ、ルーカス
ポーランド３０－１５０ クラクフ アレヤ・アルミイ・クラヨベイ１８
(72)発明者 クハルチク、シモン
ポーランド３０－１５０ クラクフ アレヤ・アルミイ・クラヨベイ１８
(72)発明者 ヒルツェル、マーティン
アメリカ合衆国１０５９８ ニューヨーク州ヨークタウン・ハイツ キッチャワン・ロード１１０１
(72)発明者 ラジャク、ドロタ
ポーランド３０－１５０ クラクフ アレヤ・アルミイ・クラヨベイ１８