FUJITSU

AI倫理影響評価適用例

改版履歴

版数	日付	内容
v1.0	2022年2月14日	新規作成

はじめに

- 本書は，AIシステムの利用で起こり得る倫理上のリスクを評価する方式として当社が開発した，AI倫理影響評価を事例に適用して示したものです．
- 昨今，AIが起こす倫理的な問題が社会で認知されるようになっています．例えば，顔認識AIが人種差別的な結果を出したり[1]，人材採用AIが性差別的な結果を出し運用停止となったというニュース[2]は大きく報道されました．一方，欧州をはじめ各国や組織はAIを活用するための倫理原則やガイドラインを策定[3][4][5][6][7][8]し，倫理的な問題に対処しようとしています．欧州委員会はさらに踏み込んだAI規則案[9]を提案しています．
- このように，AIを社会実装する上で，倫理上のリスクに対処することが必要不可欠となっている一方で，AIシステムが複数のステークホルダーを持ち，それらを取り巻く社会状況が変化することから，AIシステムの利用によりどのような倫理的な問題が発生するかを適切なタイミングで認識することは課題となっています．
- このような背景から，我々は，AIシステムの開発者や提供者が，自身のユースケースから倫理的な影響を評価するためのAI倫理影響評価を開発しました．本方式を，すでに知られている倫理的なインシデント事例を参考にして構築した事例に適用し，倫理的なリスクがAIシステムのどこにどのように起こり得るかを把握できることを確認しました．本書は，これらの適用結果を示したものです．
- 本書が，AI開発者，AIサービスプロバイダやビジネス利用者などAIに関わる皆さんにとって，倫理的な問題が起こり得ることを認識する一助となれば幸いです．
- なお，AI倫理影響評価は，個別のAIのユースケースについて，倫理的に問題ないことを保証するものではありません．
- 当社では，本書をもとに多様な立場の関係者と議論や検討を行い，本方式の改良を進めていきます．

利用方法

■ **本書の想定読者**

- AIシステムの顧客向け提案を検討している営業担当者やエンジニア
- AIアルゴリズム開発者・研究者・データサイエンティスト
- AIシステムの自社ビジネス適用を検討する企業・組織
- AIシステムを利用する人や，その利害関係者

■ **本書が対象とするAIシステムのライフサイクル**

- AIシステムの企画，設計・開発，運用までのプロセスを対象とします.

■ **本書の活用シーン**

- AIシステムの企画時に，営業担当者から顧客に，起こり得るAI倫理リスクを評価し説明する
- AIシステムの設計段階で，AIシステムの構成要素に起こり得るAI倫理リスクを評価する
- AIシステムの運用に先立ち，起こり得るAI倫理リスクを評価する
- すでに稼働しているAIシステムに対して，起こり得るAI倫理リスクを評価する
- AIシステムを利用する人が，利用前に，起こり得るAI倫理リスクを評価する

免責事項

- 本書で扱う事例は，過去に発生した事案（国際的AIコンソーシアムであるPartnership on AI[10]が公開しているAI Incident Database[11]に登録された情報）をベースに，類似事例やISOが公開しているAIユースケース[12]などを参考に構築した架空のものです．実際の事案とは異なります．
- 倫理の考え方は，個人の価値観や国・地域ごとの文化・宗教や社会状況，技術状況などによって，またその時代によっても変化します．本書の内容は富士通としての見解ではなく，あくまで例示に過ぎません．
- 富士通は，本書で指摘した以外にリスクがないことを保証するものではなく，具体的な事案における対応は，本書の読者である各団体・個人の責任において判断し，実施していただく必要があります．
- 本方式や本書に関連して読者にいかなる損害が生じた場合であっても，富士通は責任を負いません．

目次

- AI倫理影響評価とは
- 適用結果の見方と用語
- 適用例

AI倫理影響評価とは

本書におけるリスクの定義

本書では倫理的なリスクを扱います．分析では「リスク事象」「リスク要因」という言葉を用います．リスクに関する用語の説明と例は以下の通りです．

用語	説明	例
AI倫理リスク	AIシステムが引き起こす倫理的な問題に起因するリスク．AIシステムやその関係者に悪影響があることだけでなく，良い影響が有る場合も含む	ローン審査AIによって融資可と判定される人が，特定の人種や性別に偏ってしまう不公正が起こること
リスク事象	AI倫理リスクのうち，ステークホルダーに影響を与えるリスク，あるいはステークホルダーが影響を及ぼすことによるリスクのこと．リスク事象によって，ステークホルダーに経済的な損失や社会的信頼性失墜，あるいは逆に収益や社会的信頼性の増加が発生する	ローン審査AIによる審査結果について，特定の人種グループにおける融資可と判断される割合が他の人種グループに比べて極端に低く，不公正であること
リスク要因	AI倫理リスクのうち，リスク事象を引き起こす要因． リスク事象が他のリスク事象の要因となることもある	ローン審査AIの構築時に，融資判断を行うAIを作成するためのデータが，人種や性別による差別を含んでいること

AI倫理影響評価とは

AI倫理影響評価とは

- AI倫理影響評価とは，AI倫理ガイドラインに基づいて，AIシステムに発生する可能性がある倫理的なリスクを系統的に分析するための方式です．
- 本方式はAI倫理ガイドラインを具体化してチェック項目化し，AIシステムを分析して倫理的に問題が起きそうなところをリスクとして洗い出します．
- AI倫理ガイドラインとして，欧州 AI HLEGが定める倫理ガイドライン（Trustworthy AI）[4]をベースとしています．
- 本方式は，AI倫理ガイドラインの内容の範囲で，ガイドラインへのコンプライアンスをチェックするものです．ガイドラインには現実社会の倫理が入っているわけではありません．また，ガイドラインは，法で規制されているかを判断するものではありません．
- その他，本書のご利用にあたっての注意点については，5ページ「免責事項」もご覧ください．

AI倫理影響評価の全体図を図1に示します．

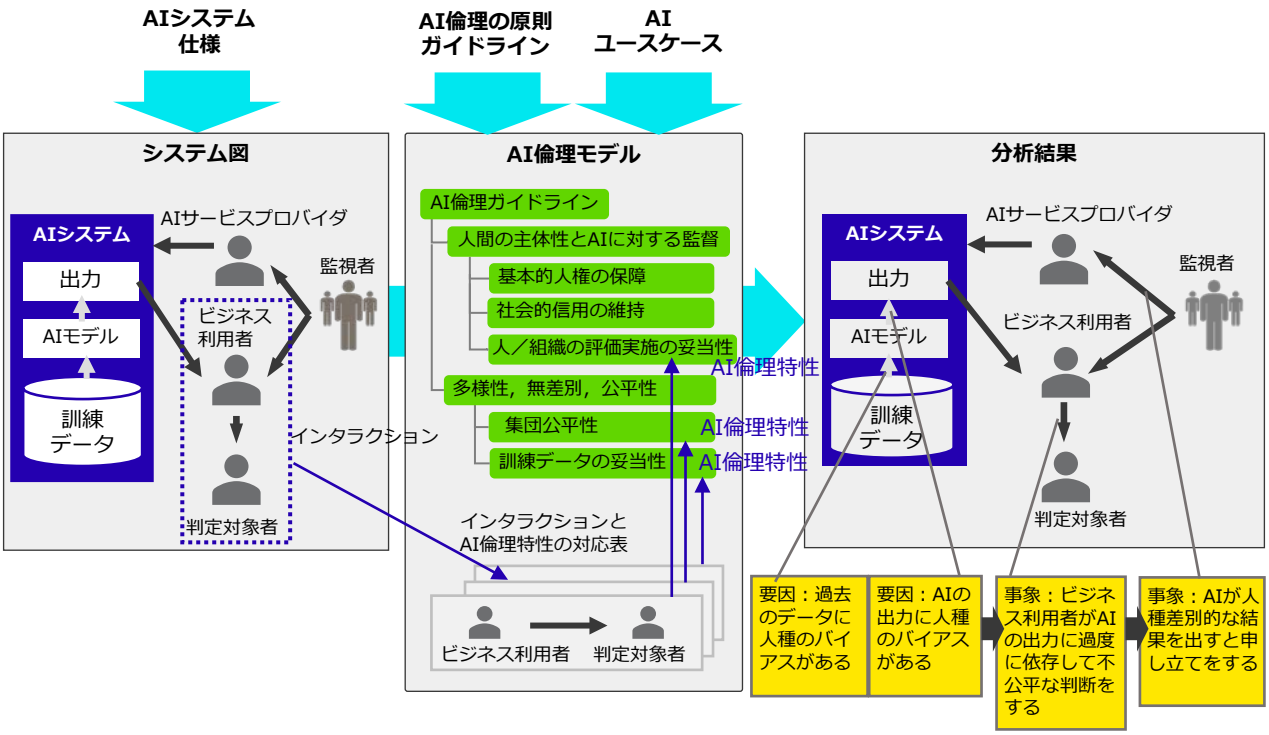


図1 AI倫理影響評価の全体図

AI倫理影響評価とは

■ システム図

- AI倫理影響評価では，まずシステム図を作成します．システム図は，AIシステムの構成要素およびステークホルダーと，ステークホルダー同士，およびAIシステムの構成要素同士の関係を洗い出し，記載したものです．この関係をインタラクションと呼びます．
- システム図は，本方式を用いて評価する人が，評価対象のAIシステムに関する仕様やユースケースの情報に基づいて作成します．

■ AI倫理モデル

- AI倫理モデルとは，AI倫理ガイドラインを構造化しAIユースケースと突き合わせて具体化したものです．AI倫理ガイドラインとして，欧州AI HLEGが提案する”Ethics Guidelines for Trustworthy AI”（以下，Trustworthy AI）に基づいて作成しています．
- AI倫理モデルは，インタラクションの種類に応じて，そのインタラクションに対応するAI倫理特性（倫理的なAIシステムが満たすべき特性）を示した対応表を提供します．
- システム図に洗い出したインタラクション毎に，対応表を用いて，起こり得るAI倫理リスクを抽出します．この抽出作業は，分析者が，分析対象のAIシステムのユースケースに基づいて，該当するAI倫理特性に反する状態を具体的なリスクのシーンとして想定する作業です．

■ 分析結果

- 抽出したリスクを整理してシステム図に記載したものです．リスクをステークホルダーとのインタラクションに現れるリスク事象と，そのリスク事象を引き起こすリスク要因とに分けて関連付けて示します．

AI倫理影響評価とは

AI倫理影響評価の用語

用語	説明
システム図	AIシステムの構成要素とステークホルダーを図示し，さらにインタラクションを記載した図
インタラクション	AIシステムの構成要素またはステークホルダーのいずれか2者間の関係．AI倫理影響評価では，倫理的なリスクをインタラクションに紐づけて抽出する
AI倫理モデル	AI倫理の原則やガイドラインをユースケースと付き合わせて具体化し，インタラクションと対応づけたもの
AI倫理特性	倫理的なAIシステムが満たすべき特性．AI倫理ガイドラインを具体化して抽出したもの． 例：「基本的人権の保障」「集団公平性」

適用結果の見方と用語

適用結果の見方

本書は，事例毎に以下のような構成で説明します．

1. **AIシステムの概要**
   AIシステムの利用シーンと構成，および想定される倫理的な問題を説明します．
2. **ユースケース概要**
   AI倫理影響評価に必要となるAIユースケースの情報を記載した表です．
3. **分析図**
   ユースケース概要から作成したシステム図に，AI倫理影響評価で抽出したAI倫理リスクを，インタラクションと対応付けて示した図です．
4. **リスクの整理**
   AI倫理影響評価で抽出した倫理的なリスクを，リスク事象とそれを引き起こすリスク要因に対応付けて整理した表です．

次頁以降にユースケース概要，分析図，リスクの整理について説明します．

適用結果の見方と用語

ユースケース概要の見方

- ユースケース概要は，AI倫理影響評価に必要となるAIユースケースの情報を記載した表です．
- AI倫理影響評価に必要な項目を「大項目」と「中項目」で示しています．各項目の「内容」カラムにユースケース毎の説明が記載されます．
- ユースケース概要の各項目の説明と記載例を表で説明します．
    - 大項目「ステークホルダーと役割」に該当する中項目に「※」が付いている項目は，総務省「AI利活用ガイドライン」[7]に定義された名称に準拠しています．

大項目	中項目	内容	項目の説明	例
業種			AI倫理事例で想定される業務や業種． 国際標準産業分類[13]に基づいて記入する	・金融・保険業 ・公務及び国防・義務的社会保障事業
目的			AIシステムの利用目的	・ローン審査を短時間で実施
サービス	サービス概要		AIシステムを用いAIシステム提供者が行うサービス	・ローン申込者にAIが融資判断を回答する
	顧客毎のカスタマイズの有無		顧客ごとにカスタマイズが必要なシステムであるか否かを示す	有り／無し／不明のいずれかを記載
	要件		顧客からサービスに対して求められている要件	有り／無し／不明のいずれかを記載
利用シーン			AIシステムの利用者や利用環境の特徴，AIシステムを使って行う作業の概要を示す	・ローン申込者はアプリで申込情報を送信，審査結果をアプリで受け取る
ステークホルダーと役割	AIサービスプロバイダ※		ステークホルダの一種．開発されたAIシステムを用いて業務を行い，各種サービスを提供する	・ローン審査AIサービス開発ベンダ
	開発者※		ステークホルダの一種．AIシステムを開発する	・AI開発ベンダ
	ビジネス利用者※		ステークホルダの一種．業としてAIシステム又はAIサービスを利用する者	・ローン審査AIの推論結果から融資判断をする銀行の担当者 ・顔認識AIが不審人物と認識した人物について捜査判断を決定する警察官 ・医療画像診断AIの推論結果に基づいて治療方針を決定する医師
	消費者的利用者※		ステークホルダの一種． ビジネス利用者を除く，AIシステム又はAIサービスを利用する者	・ローン審査AIの申込者 ・医療画像診断を受ける患者 ・チャットボットのユーザ

適用結果の見方と用語

大項目	中項目	内容	項目の説明	例
ステークホルダーと役割	訓練データ提供者		ステークホルダの一種．訓練データを作成するための元データを提供する人	・信用調査機関，銀行 ・自然言語処理用のデータセットの保有者
	訓練データ取得元		ステークホルダの一種．訓練データ提供者と直接／間接に関わりがある人	・信用調査機関，銀行 ・ 顔画像データセットに自分の顔画像を提供する人
	訓練データ取得時の関係者		ステークホルダの一種．訓練データ取得をする際に直接／間接的に関わりがある人，組織，システム	・信用スコアを設計する人，過去に信用スコアが提供された人達 ・顔画像を撮影するカメラマン ・医療画像を撮影する医師や技師
	推論データ提供者		ステークホルダの一種．推論データを作成するため入力データを提供する人	・信用調査機関，銀行 ・医療画像データを提供する病院
	推論データ取得元		ステークホルダの一種．推論データの内容に関わりがある人や組織	・信用調査機関，銀行 ・医療画像データに自身の画像を使用される患者
	推論データ取得時の関係者		ステークホルダの一種．推論データ取得をする際に直接／間接的に関わりがある人	・顔画像を撮影するカメラマン ・医療画像を撮影する医師や技師 ・画像に映り込んでいる人
	監視者		ステークホルダの一種． AIシステム又はAIサービスを監視する人	・人権団体，メディア
	サービスUI/API提供者		AIシステムの推論結果をもとに，AIシステム利用者向けの機能を構築する人や組織，システム	・医療画像診断のモニタ・操作部の提供者 ・工場ロボットのロボット制御部の提供者 ・チャットボットが動作するSNSなどのプラットフォーム提供者
	判定対象者		AIシステムによって何らかの判定，評価をされる人や組織	・監視カメラに写り込んだ人 ・人事AIで評価される従業員
	サービス認可者		AIシステムの開発やAIシステムを用いて行うサービス提供を認可する人や組織	・関係省庁，規制当局
	その他のステークホルダー		ステークホルダの一種．AIシステムからの出力やAIサービスに間接的に影響を受ける人や組織	・利用者の家族やビジネスパートナー ・利用者が契約する保険会社 ・利用者の属するコミュニティー

適用結果の見方と用語

大項目	中項目	内容	項目の説明	例
Human-in-the-loopの有無			AIシステムで，推論結果に対してAIシステム利用者の判断が加わるか否かを示す．人の判断が加わるか否かで抽出するリスク事象やリスク要因が異なる	・銀行の融資担当者が融資可／不可の最終判断をする場合はhuman-in-the-loop有．AIの推論結果を直接ローン申込者に伝える場合はhuman-in-the-loop無し
既存手段の有無			AIシステムで行う／支援するタスクが，既存の手段で行えるか否かを表すもの．既存手段で行える場合は，その既存手段とAIシステムとで効率性や精度の比較が必要	
AIタスク	タスク		AIモデルへの入力である推論データと出力である推論結果の概要を示したもの	・ローン申請者の申請情報，信用スコアおよび取引データを入力し，融資判断を出力する
	問題種別		AIモデルで扱う問題の種類	分類，推薦，回帰，自然言語処理，音声認識，画像認識，画像生成
	出力		AIモデルの出力	・融資の可・不可 ・チャットボットの会話文
	技術		AIモデルの入手先．AIシステム開発時に作成したものか，他者が作成したものであるか，あるいは組み合わせたものであるかを示す	自前，OSS，自前＋OSS，他社技術
	AIタスクの実施に必要なモデルの内訳		AIタスクの実施に必要なすべてのAIモデルの名称を示す	・音声認識モデル，文章理解モデル，表情認識モデル
	学習データの更新・追加学習の有無		運用開始後に，更新した訓練データで追加学習が必要であるかを示す	有り／無し／不明のいずれかを記載
	リアルタイム性		AIタスクの処理に，リアルタイム性が求められるか否か	有り／無し／不明のいずれかを記載
	AIタスクの検証（既存手段による判断との整合性）		既存の手段で同様のタスクを行っていた場合，その既存手段と比較してAIタスクとしての精度あるいは個々の結果の相違点を示すもの	・ローン審査の担当者による判断との比較

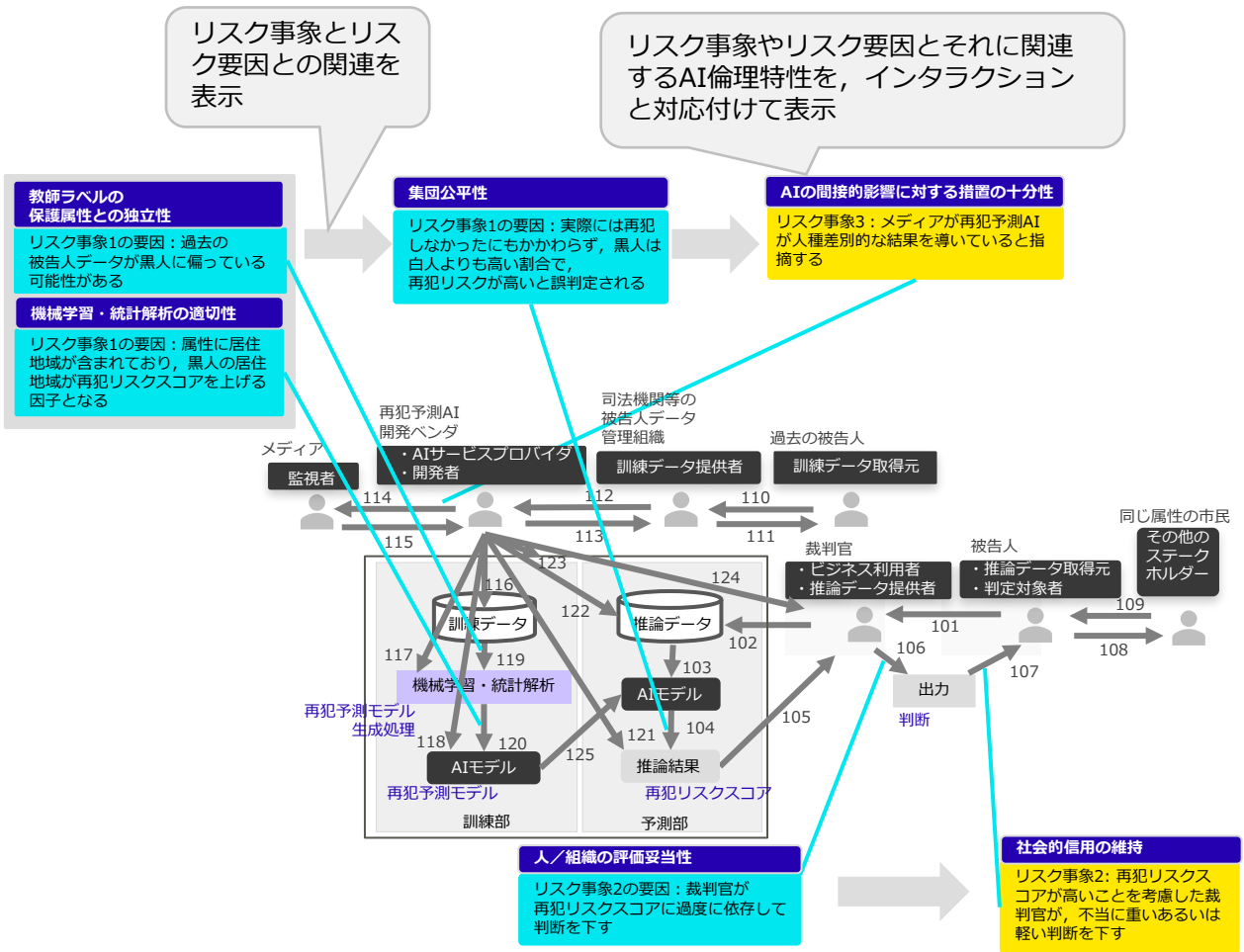
適用結果の見方と用語

大項目	中項目	内容	項目の説明	例
訓練データ	内訳と提供者および取得元		訓練データの内訳と，それぞれのデータ提供者とデータ取得元	・過去のローン申込者の取引データ（銀行）
	教師ラベル		訓練データに設定される教師ラベルの内容	・融資可/不可
	保護属性，公平性検証・緩和に使う属性		訓練データに含まれる保護属性（性別，年齢，国籍，人種）	・性別，年齢
	個人情報の有無		訓練データに個人情報が含まれるかを示すもの	有り／無し／不明のいずれかを記入
	データ取得時の留意事項		訓練データを取得する際に気を付けること	訓練データに関わる人に許可を得る
	データの保存		訓練データの保存可否や保存の際の注意事項	訓練データの保存は不可
推論データ	内訳と提供者および取得元		推論データの内訳と，それぞれのデータ提供者とデータ取得元	・ローン申込者の信用スコア（信用調査機関）
	保護属性，公平性検証・緩和に使う属性		推論データに含まれる保護属性（性別，年齢，国籍，人種）	・性別，年齢
	個人情報の有無		推論データに個人情報が含まれるかを示すもの	有り／無し／不明のいずれかを記入
	データ取得時の留意事項		推論データを取得する際に気を付けること	データ取得元の許諾が必要か．許諾条件は何かなど
AIシステムの出力	出力時の留意事項		AIシステムの出力時に気を付けること	・AIの結果についての相談窓口を提示する
引用元，参考記事			インシデントデータベースの登録URLや解説記事へのリンク	

適用結果の見方と用語

分析図の見方

- システム図に，AI倫理影響評価によって抽出されたリスク事象およびリスク要因を，インタラクションと対応付けて示したもの



分析図中の記号の説明

	ステークホルダー
	AIシステム
	インタラクション(数字はインタラクションID)

	リスク事象
	リスク要因
	AI倫理特性
	リスク事象あるいはリスク要因とインタラクションとの対応

適用結果の見方と用語

リスクの整理の見方

- AI倫理影響評価によって抽出したリスク事象とそれを引き起こしたリスク要因を対応付けて整理した表

リスク事象とそれを引き起こすリスク要因を並べて記載

リスク要因が空欄の行はリスク事象と対応するAI倫理特性とインタラクションIDを記載
それ以外の行はリスク要因と対応するAI倫理特性とインタラクションIDを記載

リスク事象	リスク要因	AI倫理特性	インタラクションID
①実際には再犯しなかったにもかかわらず，黒人は白人よりも高い割合で，再犯リスクが高いと誤判定される可能性がある		集団公平性	104
	属性に居住地域が含まれており，黒人の居住地域がリスクスコアを上げる因子となる	機械学習・統計解析の適切性	120
	過去の被告人データが黒人に偏っている可能性がある	教師ラベルの保護属性との独立性	119
②再犯リスクスコアが高いことを考慮した裁判官が，不当に重いあるいは軽い判断を下す		社会的信用の維持	107
	裁判官が再犯リスクスコアに過度に依存して判断を下す	人／組織の評価実施の妥当性	106
③メディアが再犯予測AIの結果が人種差別を含むと指摘する		AIの間接的影響に対する措置の十分性	114
	リスク事象①	集団公平性	104

リスク事象の他のリスク事象の要因となる場合は，リスク事象に番号を振り，その番号をリスク要因に記載

適用例

適用例一覧

No.	名称	業種（国際標準産業分類[13]）	AIタスク	データ種類	想定される倫理的な問題
1	チャットボット	芸術・娯楽及びレクリエーション	文章理解・生成	自然言語	差別的な表現を発信する.
2	採用AI	管理・支援サービス業	分類	テーブルデータ	採用結果に女性差別がある.
3	再犯リスク予測	公務及び国防・義務的社会保障事業	分類	テーブルデータ	AIの予測結果が，黒人に対して不公平である.
4	警察による顔認識	公務及び国防・義務的社会保障事業	分類	画像	AIが監視カメラに映った容疑者の顔画像を，市民の顔データベースと照合し，無関係の市民が誤認逮捕される.
5	ビデオ面接の採用判定AI	管理・支援サービス業	分類	画像	表情認識AIによる判定結果が公平でない可能性がある.
6	ローン審査AI	金融・保険業	分類	テーブルデータ	女性や若年事業者は，年配の男性事業者に比べて審査に通りにくい.
7	果物等級判定	農業・林業及び漁業	分類	画像	目的外である農家の評価に使用する.

1. チャットボット

AIシステムの概要

■ AIシステムの利用シーン
- SNSの不特定ユーザと会話をするチャットボット

■ AIシステムの構成
- SNSのユーザからの問いかけに対してAIが返答する
- 文章理解・生成モデル
- 不特定ユーザとの会話を学習する

■ 想定される倫理的な問題
- チャットボットが悪意のあるユーザの会話を学習し，人種差別的，性差別的，暴力的な会話を発信する．

ユースケース概要

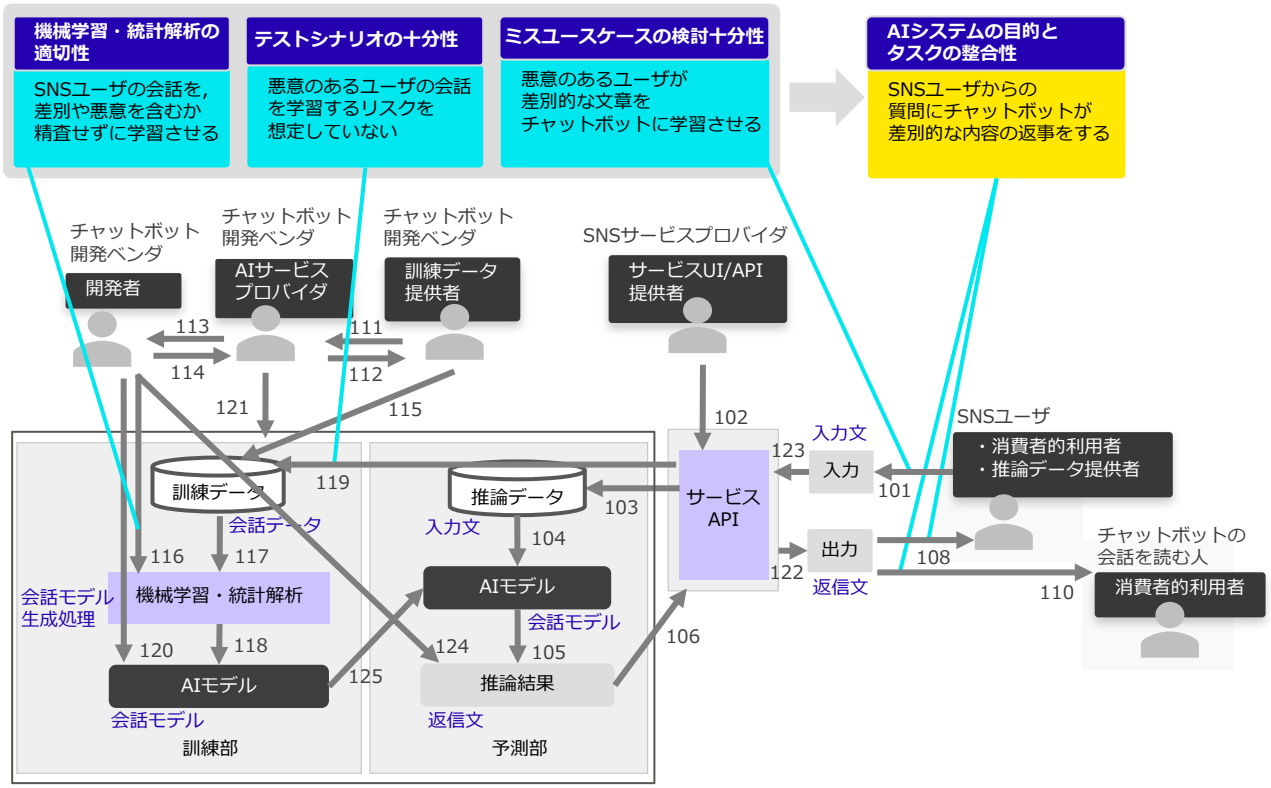
大項目	中項目	内容
業種		エンターテイメント
目的		ユーザと自由な会話を学習するチャットボット
サービス	サービス概要	SNS上でユーザが話しかけた内容に意味のある返事をするチャットボット
	顧客毎のカスタマイズの有無	無し
	要件	無し
利用シーン		ユーザがチャットボットと会話する
ステークホルダーと役割	AIサービスプロバイダ	チャットボット開発ベンダ
	開発者	チャットボット開発ベンダ
	ビジネス利用者	無し
	訓練データ提供者	チャットボット開発ベンダ
	訓練データ取得元	チャットボット開発ベンダ
	訓練データ取得時の関係者	不明
	推論データ提供者	SNSユーザ
	推論データ取得元	SNSユーザ
	推論データ取得時の関係者	不明
	消費者的利用者	SNSユーザ
	監視者	無し
	サービスUI/API提供者	SNSサービスプロバイダ
	判定対象者	無し
	サービス認可者	無し

1. チャットボット

大項目	中項目	内容
ステークホルダーと役割	その他のステークホルダー	チャットボットの会話を見る人やメディアなどの組織
Human-in-the-loopの有無		無し
既存手段の有無		無し
AIタスク	タスク	文章理解・生成
	問題種別	自然言語処理
	出力ラベル	文章
	技術	不明
	AIタスクの実施に必要なモデルの内訳	文章理解・生成モデル
	学習データの更新・追加学習の有無	有り
	リアルタイム性	有り
	AIタスクの検証（既存手段による判断との整合性）	無し
訓練データ	内訳と提供者および取得元	SNSユーザの入力文（提供者と取得元：開発ベンダ，SNSユーザ）
	教師ラベル	無し
	保護属性，公平性検証・緩和に使う属性	無し
	個人情報の有無	無し
	データ取得時の留意事項	不明
	データの保存	無し
推論データ	内訳とデーターオーナー	SNSユーザの入力文（提供者と取得元：SNSユーザ）
	保護属性，公平性検証・緩和に使う属性	無し
	個人情報の有無	無し
	データ取得時の留意事項	無し
AIシステムの出力	出力時の留意事項	無し
引用元，参考記事		PAI AI incident database: incident ID #6 https://incidentdatabase.ai/cite/6#undefined

1. チャットボット

分析図



リスクの整理

リスク事象	リスク要因	AI倫理特性	インタラクションID
SNSユーザからの質問にチャットボットが差別的な内容の返事をする		AIシステムの目的とタスクの整合性	108, 110
	差別や悪意を含む会話を学習する	機械学習・統計解析の適切性	116
	悪意のあるユーザの会話を学習するリスクを想定していない	テストシナリオの十分性	119
	悪意のあるユーザが差別的な文章をチャットボットに学習させる	ミスユースケースの検討十分性	101

2. 採用AI

AIシステムの概要

■ AIシステムの利用シーン

- 人材採用において，求職者の履歴書から採用面接の対象者をスクリーニングする

■ AIシステムの構成

- 過去の求職者の履歴書と採用結果を訓練し，採用判定AIを生成．人種および性別は履歴書に記載されない．

■ 想定される倫理的な問題

- AIによる採用候補者が極端に男性に偏る

ユースケース概要

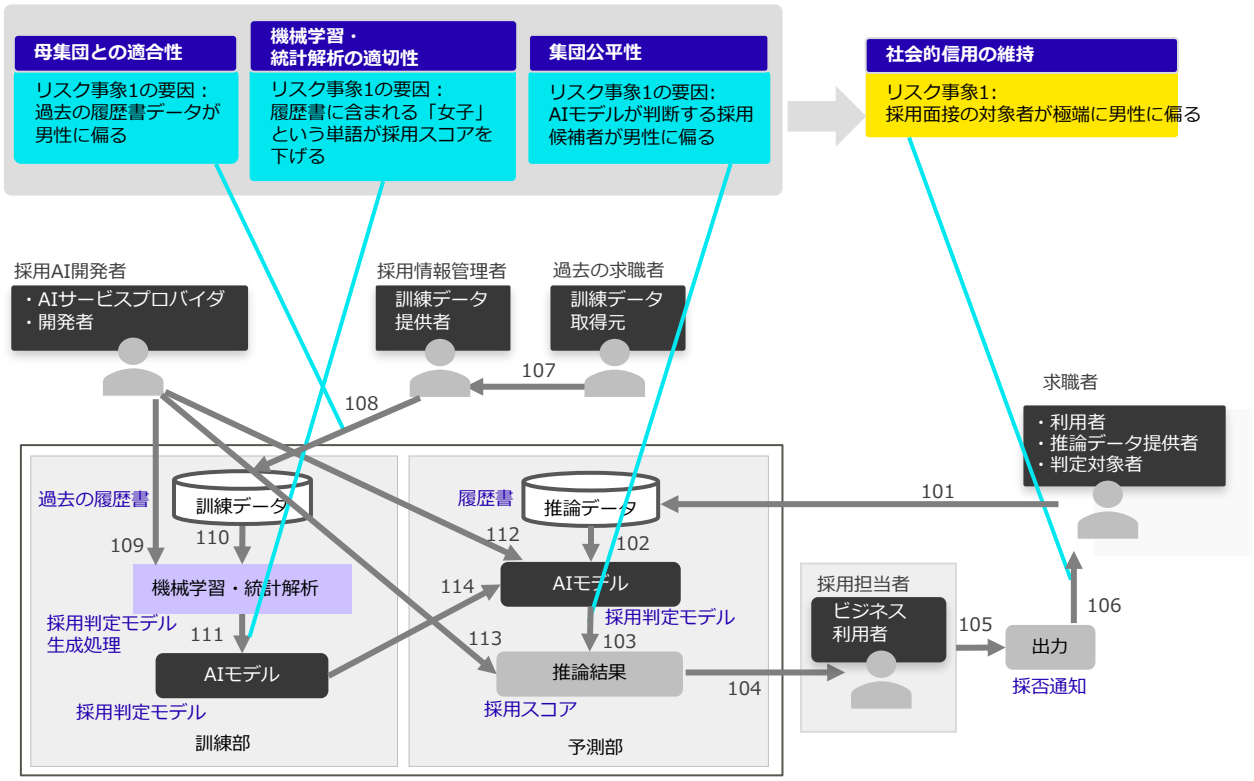
大項目	中項目	内容
業種		管理・支援サービス業
目的		採用面接候補者の絞り込み
サービス	サービス概要	自社の求職者の履歴書を自動的に分類し，最も有望な候補を選択するAIツール
	顧客毎のカスタマイズの有無	無し
	要件	無し
利用シーン		自社の求職者の履歴書を自動的に分類し，最も有望な候補を選択する
ステークホルダーと役割	AIサービスプロバイダ	採用AI開発ベンダ
	開発者	採用AI開発ベンダ
	ビジネス利用者	採用AIのユーザ企業の採用担当
	訓練データ提供者	採用AIのユーザ企業
	訓練データ取得元	採用AIのユーザ企業
	訓練データ取得時の関係者	過去の求職者
	推論データ提供者	求職者
	推論データ取得元	求職者
	推論データ取得時の関係者	不明
	消費者的利用者	求職者
	監視者	不明
	サービスUI/API提供者	求職者
	判定対象者	求職者
	サービス認可者	不明
	その他のステークホルダー	不明

2. 採用AI

大項目	中項目	内容
Human-in-the-loopの有無		無し
既存手段の有無		履歴書の内容から求人応募者をスコアリング
AIタスク	タスク	履歴書データから採用候補者のレベルを判定する
	問題種別	分類問題
	出力ラベル	5段階のスコア
	技術	無し
	AIタスクの実施に必要なモデルの内訳	無し
	学習データの更新・追加学習の有無	無し
	リアルタイム性	無し
	AIタスクの検証（既存手段による判断との整合性）	採用担当者による判断
訓練データ	内訳と提供者および取得元	過去の履歴書(提供者：採用AIのユーザ企業，取得元：過去の求職者)
	教師ラベル	5段階のスコア
	保護属性，公平性検証・緩和に使う属性	性別
	個人情報の有無	不明
	データ取得時の留意事項	不明
	データの保存	不明
推論データ	内訳と提供者および取得元	履歴書（提供者と取得元：求職者）
	保護属性，公平性検証・緩和に使う属性	性別
	個人情報の有無	有り
	データ取得時の留意事項	不明
AIシステムの出力	出力時の留意事項	不明
引用元，参考記事		PAI incident database #37 https://incidentdatabase.ai/cite/37

2. 採用AI

分析図



リスクの整理

リスク事象	リスク要因	AI倫理特性	インタラクションID
採用面接の対象者が極端に男性に偏る		社会的信用の維持	106
	過去の履歴書データの採用候補者が男性に偏る	母集団との適合性	108
	履歴書に含まれる「女子」という単語が採用スコアを下げる	機械学習・統計解析の適切性	111
	AIが判断する採用候補者が男性に偏り，女子大の卒業生は採用スコアが低く判定される	集団公平性	103

3. 再犯リスク予測

AIシステムの概要

■ AIシステムの利用シーン

- 再犯リスク予測AI: 被告人に関する情報(*)を入力データとして，AIが被告人の再犯リスクを10段階のスコアで判定する．裁判官は被告人の仮釈放の有無や量刑の判断を下す際に，再犯リスクスコアを参考にする．
  (*)被告人の犯罪歴，薬物使用の有無，教育や雇用レベルを含む．人種は含まない．

■ 想定される倫理的な問題

- 再犯リスク予測AIが人種差別的な結果を導く

※各国の刑事・司法などの法制度については，本書における評価の対象外です．

ユースケース概要

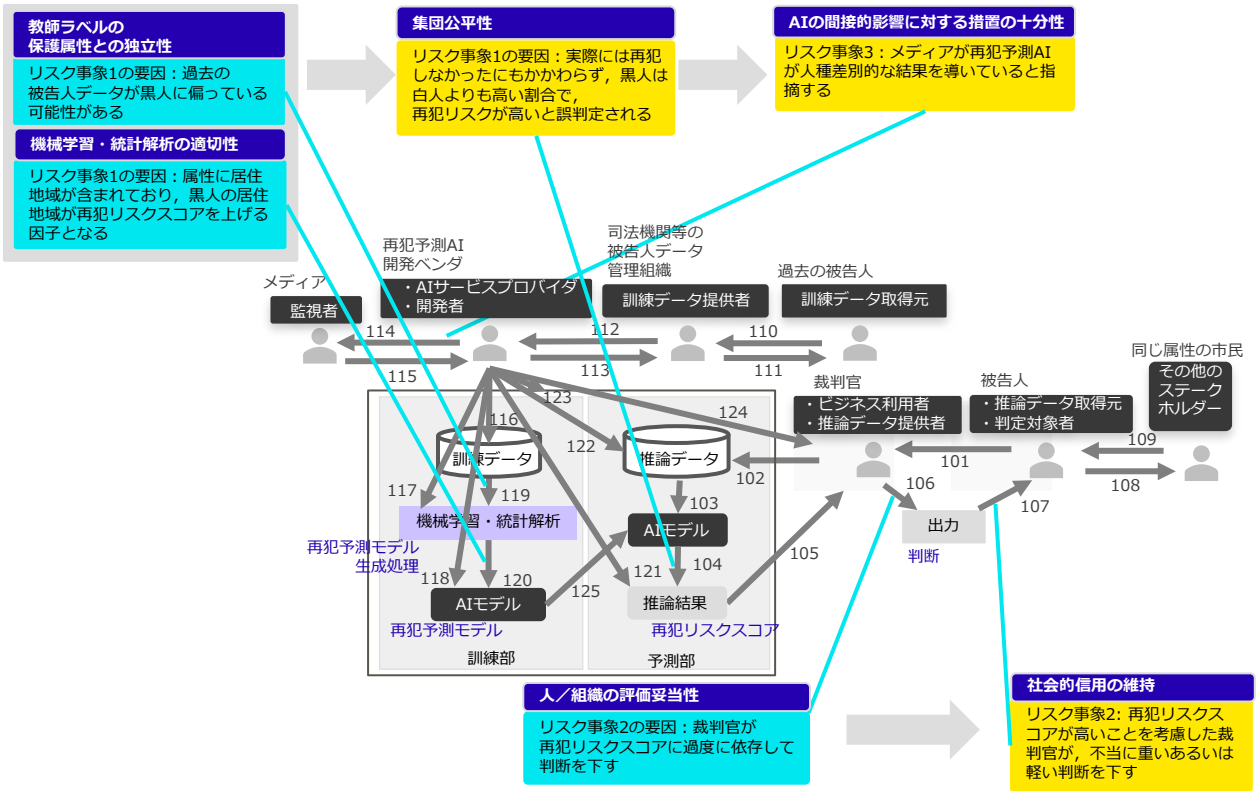
大項目	中項目	内容
業種		公務及び国防・義務的社会保障事業
目的		裁判官が仮釈放と判決に関する決定をするための参考利用
サービス	サービス概要	被告人の再犯リスクを10段階評価で判定する
	顧客毎のカスタマイズの有無	無し
	要件	不明
利用シーン		裁判官が裁判前の仮釈放と再犯の判決に関する決定のために再犯リスクスコアを参考にする
ステークホルダーと役割	AIサービスプロバイダ	再犯リスク予測AI開発ベンダ
	開発者	再犯リスク予測AI開発ベンダ
	ビジネス利用者	裁判所の裁判官
	訓練データ提供者	被告人データを管理する組織
	訓練データ取得元	過去の被告人
	訓練データ取得時の関係者	不明
	推論データ提供者	裁判所
	推論データ取得元	被告人
	推論データ取得時の関係者	不明
	消費者的利用者	無し
	監視者	メディア，一般市民（メディアのニュースを読んだ人，興味をもってデータ分析した人など）
	サービスUI/API提供者	無し
	判定対象者	被告人

3. 再犯リスク予測

大項目	中項目	内容
ステークホルダーと役割	サービス認可者	無し
	その他のステークホルダー	被告と同じ属性を持つ人
Human-in-the-loopの有無		無し
既存手段の有無		裁判官による判断
AIタスク	タスク	被告人の再犯リスクを10段階評価で推定する
	問題種別	分類問題
	出力	10段階の再犯リスクスコア
	技術	不明
	AIタスクの実施に必要なモデルの内訳	再犯リスク予測モデル
	学習データの更新・追加学習の有無	不明
	リアルタイム性	不明
	AIタスクの検証（既存手段による判断との整合性）	過去の判例
訓練データ	内訳と提供者および取得元	過去の被告人に関する情報．犯罪歴，薬物使用の有無，教育や雇用レベルを含む．人種は含まない．（提供者と取得元：当局）
	教師ラベル	10段階の再犯リスクスコア
	保護属性，公平性検証・緩和に使う属性	年齢，性別
	個人情報の有無	不明
	データ取得時の留意事項	不明
	データの保存	不明
推論データ	内訳と提供者および取得元	被告人に関する情報（人種は含まない）（提供者：裁判所，取得元：被告人）
	保護属性，公平性検証・緩和に使う属性	年齢，性別
	個人情報の有無	有り
	データ取得時の留意事項	不明
AIシステムの出力	出力時の留意事項	不明
引用元，参考記事		PAI Artificial Intelligence Incident Database #11 https://incidentdatabase.ai/cite/11/

3. 再犯リスク予測

分析図



リスクの整理

リスク事象	リスク要因	AI倫理特性	インタラクションID
①実際には再犯しなかったにもかかわらず，黒人は白人よりも高い割合で，再犯リスクが高いと誤判定される可能性がある		集団公平性	104
	属性に居住地域が含まれており，黒人の居住地域がリスクスコアを上げる因子となる	機械学習・統計解析の適切性	120
	過去の被告人データが黒人に偏っている可能性がある	教師ラベルの保護属性との独立性	119
②再犯リスクスコアが高いことを考慮した裁判官が，不当に重いあるいは軽い判断を下す		社会的信用の維持	107
	裁判官が再犯リスクスコアに過度に依存して判断を下す	人／組織の評価実施の妥当性	106
③メディアが再犯予測AIの結果が人種差別を含むと指摘する		AIの間接的影響に対する措置の十分性	114
	リスク事象①	集団公平性	104

4. 警察による顔認識

AIシステムの概要

■ AIシステムの利用シーン

- 監視カメラに映った人物画像と顔データベースを照合し，顔データベースから似た人物を抽出する

■ AIシステムの構成

- 監視カメラに映った人物画像を入力とし，顔認識AIが顔データベースと照合して似ている人物を抽出する

■ 想定される倫理的な問題

- 顔認識AIが誤って無関係の人物を容疑者と一致すると判定し，誤認逮捕につながる

※各国の刑事・司法などの法制度については，本書における評価の対象外です．

ユースケース概要

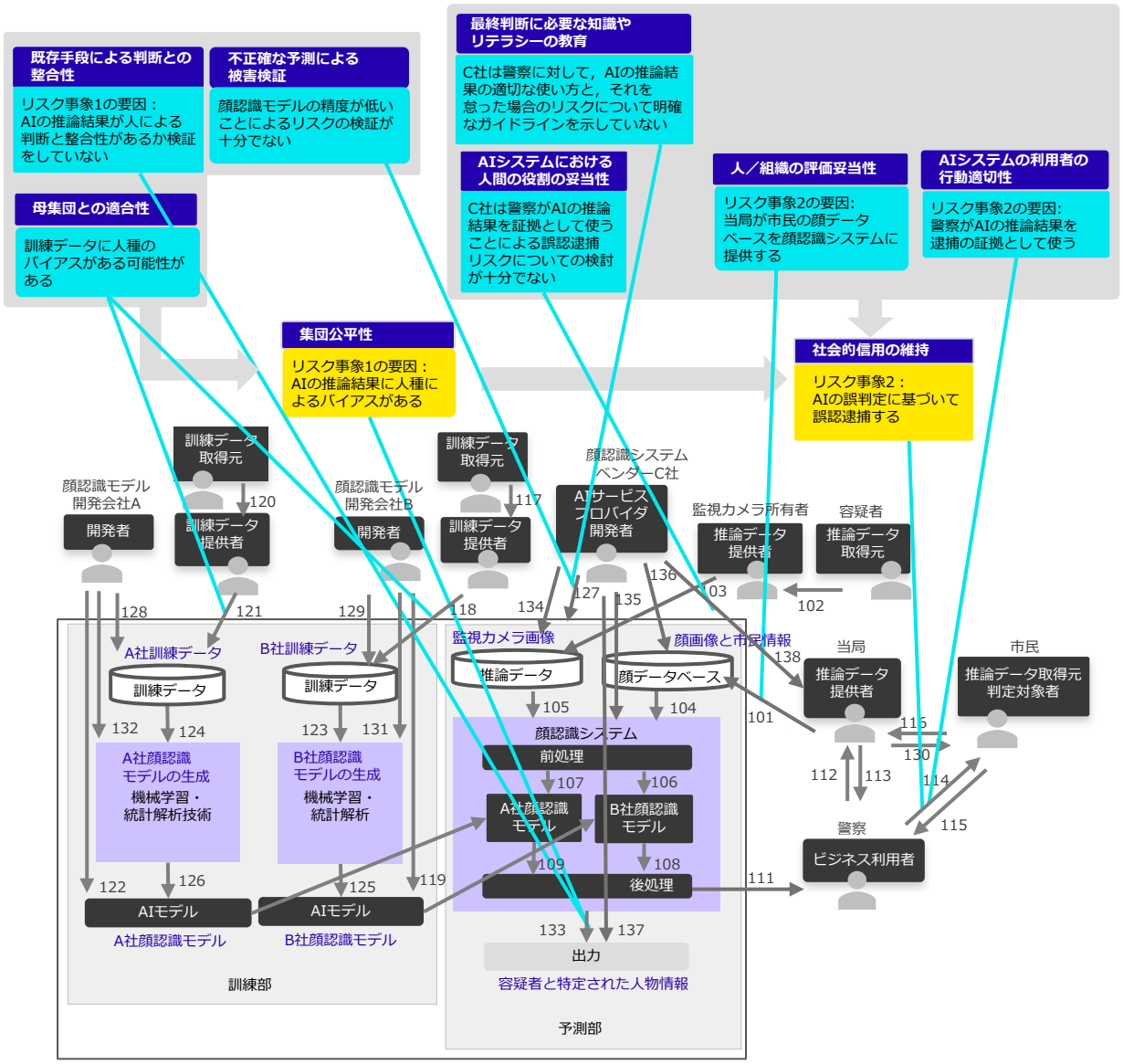
大項目	中項目	内容
業種		公務及び国防・義務的社会保障事業
目的		犯罪者の特定
サービス	サービス概要	入力された顔画像と監視カメラ画像に映った人物が一致するかを判定する
	顧客毎のカスタマイズの有無	無し
	要件	無し
利用シーン		監視カメラに映った人物画像と顔データベースを照合し，顔データベースから似た人物を抽出する
ステークホルダーと役割	AIサービスプロバイダ	顔認識システムベンダC社
	開発者	顔認識システムベンダC社，顔認識モデルベンダA社，顔認識モデルベンダB社
	ビジネス利用者	警察
	訓練データ提供者	不明
	訓練データ取得元	不明
	訓練データ取得時の関係者	不明
	推論データ提供者	①監視カメラ所有者，②顔データベース管理当局
	推論データ取得元	①監視カメラに写った容疑者，②市民
	推論データ取得時の関係者	不明
	消費者的利用者	無し
	監視者	不明
	サービスUI/API提供者	無し
	判定対象者	市民
	サービス認可者	無し
	その他のステークホルダー	市民

4. 警察による顔認識

大項目	中項目	内容
Human-in-the-loopの有無		有り（警察）
既存手段の有無		人による判断
AIタスク	タスク	入力された顔画像データから，特定したい人物の顔と類似する顔画像を検出する
	問題種別	顔認識
	出力	顔が一致するか
	技術	顔認識技術
	AIタスクの実施に必要なモデルの内訳	顔認識モデル
	学習データの更新・追加学習の有無	不明
	リアルタイム性	不明
	AIタスクの検証（既存手段による判断との整合性）	不明
訓練データ	内訳と提供者および取得元	不明
	教師ラベル	不明
	保護属性，公平性検証・緩和に使う属性	無し
	個人情報の有無	無し
	データ取得時の留意事項	不明
	データの保存時の留意事項	無し
推論データ	内訳と提供者および取得元	監視カメラの画像（提供者：監視カメラオーナー，取得元：画像に映った人物），市民の顔データベース（提供者：顔データベース管理当局，取得元：市民）
	保護属性，公平性検証・緩和に使う属性	無し
	個人情報の有無	有り
	データ取得時の留意事項	不明
AIシステムの出力	出力時の留意事項	無し
引用元，参考記事		https://incidentdatabase.ai/cite/74

4. 警察による顔認識

分析図



4. 警察による顔認識

リスクの整理

リスク事象	リスク要因	AI倫理特性	インタラクションID
①AIの推論結果に人種によるバイアスがある		集団公平性	133
	訓練データに人種のバイアスがある可能性がある	母集団との適合性	118，121
	AIの推論結果が人による判断と整合性があるか検証をしていない	既存手段による判断との整合性	133
	顔認識モデルの精度が低いことによるリスクの検証が十分でない	不正確な予測による被害検証	127
②AIの誤判定に基づいて，誤認逮捕する		社会的信用の維持	114
	リスク事象①	集団公平性	133
	AIサービスプロバイダC社は警察がAIの推論結果を証拠として使うことによる誤認逮捕リスクについての検討が十分でない	AIシステムにおける人間の役割の妥当性	127
	AIサービスプロバイダC社は警察に対して，AIの推論結果の適切な使い方と，それを怠った場合のリスクについて明確なガイドラインを示していない	最終判断に必要な知識やリテラシーの教育	138
	顔データベース管理当局が市民の顔データベースを顔認識システムに提供する	データの個人情報包含の妥当性	101
	警察がAIの推論結果を逮捕の証拠として使う	AIシステムの利用者の行動適切性	114

5. ビデオ面接の採用判定AI

AIシステムの概要

■ AIシステムの利用シーン

- ビデオ面接を使った採用判定AI：求職者のビデオ面接画像から，求職者の評価をするAIシステム
- 求職者はウェブカメラの前に座って質問に答える．AIは求職者の言葉，音声，表情から求職者の特性を評価する．

■ AIシステムの構成

- AIシステムは音声認識モデルと表情認識モデルを有する．それぞれのモデルがビデオ面接画像から判定した求職者の特性を用いて，求職者の評価結果を出力する

■ 想定される倫理的な問題

- 表情認識モデルに人種のバイアスがあるため，AIによる求職者の評価が人種差別的な結果を導く．

ユースケース概要

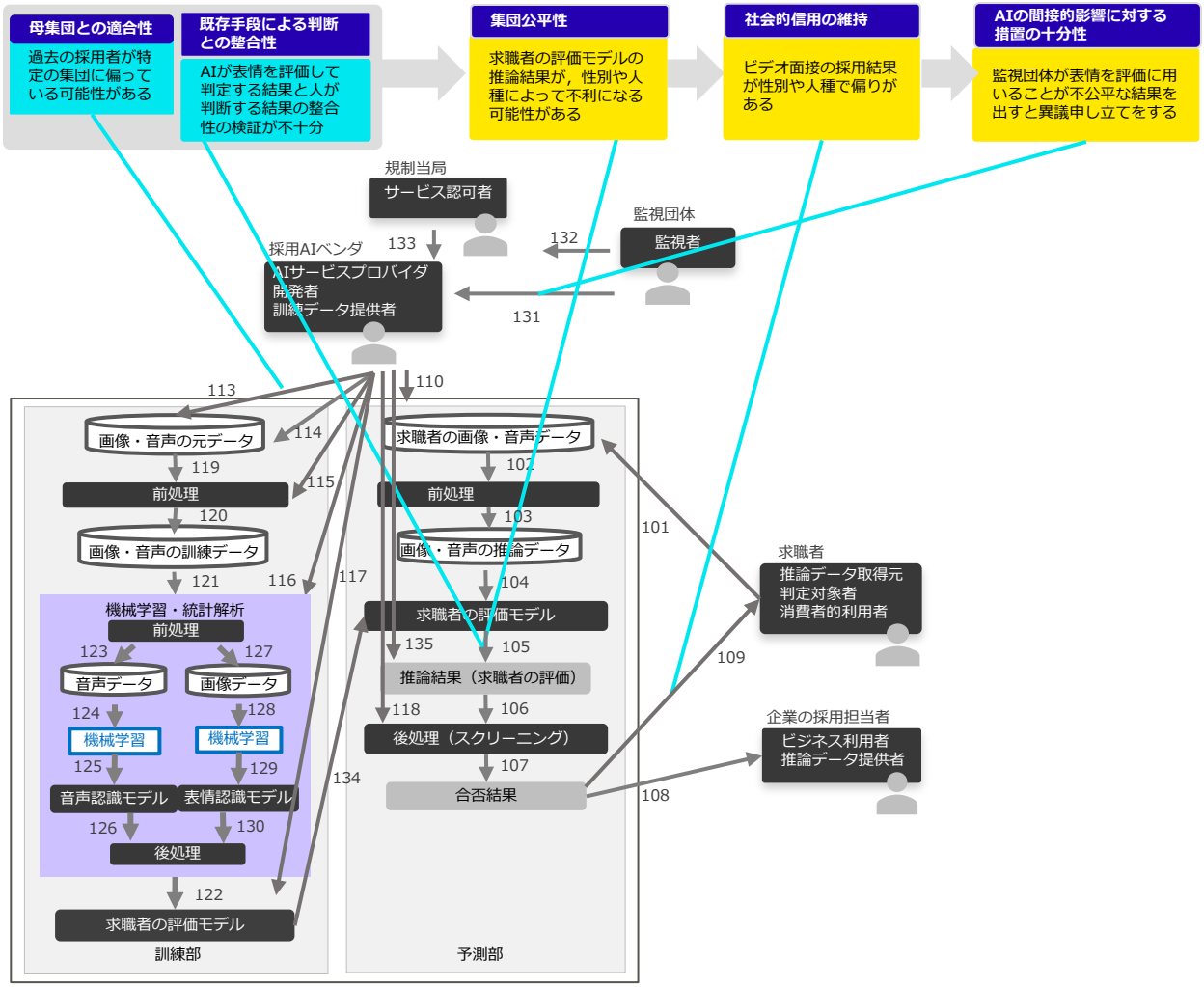
大項目	中項目	内容
業種		管理・支援サービス業
目的		求職者のビデオ面接を自動スクリーニング
サービス	サービス概要	求職者がビデオで面接の質問に答える様子をAIが評価する．求職者の言葉，音声，表情から求職者の特性を評価する
	顧客毎のカスタマイズの有無	不明
	要件	不明
利用シーン		求職者はウェブカメラの前に座って質問に答える．AIは求職者の言葉，音声，表情から求職者の特性を評価する
ステークホルダーと役割	AIサービスプロバイダ	採用AIサービスプロバイダ
	開発者	採用AIサービスプロバイダ
	ビジネス利用者	企業の人事部門
	訓練データ提供者	採用AIサービスプロバイダ
	訓練データ取得元	不明
	訓練データ取得時の関係者	不明
	推論データ提供者	求職者
	推論データ取得元	求職者
	推論データ取得時の関係者	不明
	消費者的利用者	求職者

5. ビデオ面接の採用判定AI

大項目	中項目	内容
ステークホルダーと役割	サービスUI/API提供者	採用AIサービスプロバイダ
	判定対象者	求職者
	サービス認可者	無し
	その他のステークホルダー	求職者の家族，コミュニティ
Human-in-the-loopの有無		無し
既存手段の有無		人による判断
AIタスク	タスク	求職者のビデオ面接動画から，求職者の評価を判定する
	問題種別	分類問題
	出力	求職者の評価スコア
	技術	不明
	AIタスクの実施に必要なモデルの内訳	表情分析，音声分析，自然言語処理
	学習データの更新・追加学習の有無	不明
	リアルタイム性	有り
	AIタスクの検証（既存手段による判断との整合性）	不明
訓練データ	内訳と提供者および取得元	表情データ，音声データ，会話データ(提供者と取得元：採用AIサービスプロバイダ)
	教師ラベル	人物特性評価
	保護属性，公平性検証・緩和に使う属性	無し
	個人情報の有無	無し
	データ取得時の留意事項	不明
	データの保存時の留意事項	不明
推論データ	内訳と提供者および取得元	ビデオ面接画像（提供者：ビジネス利用者の企業，取得元：求職者）
	保護属性，公平性検証・緩和に使う属性	無し
	個人情報の有無	有り
	データ取得時の留意事項	プライバシーに関する法規制を順守する
AIシステムの出力	出力時の留意事項	不明
引用元，参考記事		https://incidentdatabase.ai/cite/95

5. ビデオ面接の採用判定AI

分析図



5. ビデオ面接の採用判定AI

リスクの整理

リスク事象	リスク要因	AI倫理特性	インタラクションID
①求職者の評価モデルの推論結果が，性別や人種によって不利になる可能性がある		集団公平性	105
	AIが表情を評価して判定する結果と人が判断する結果の整合性の検証が不十分な可能性がある	既存手段による判断との整合性	105
	過去の採用者が特定の集団に偏っている可能性がある	母集団との適合性	113
②ビデオ面接の採用結果が性別や人種で偏りがある		社会的信用の維持	109
	リスク事象①	集団公平性	105
③監視団体が表情を評価に用いることが不公平な結果を出すと異議申し立てをする		AIの間接的影響に対する措置の十分性	131
	リスク事象②	社会的信用の維持	109

6. ローン審査AI

AIシステムの概要

Case1

■ AIシステムの利用シーン

- 銀行のローン審査業務において，従来融資担当者が行っていた融資判断をAIが判断する.

■ AIシステムの構成

- ローン申込者は，ローン審査AIサービスアプリに，申込者情報，融資金額および返済期間を入力する.
- ローン審査AIは入力された情報を元に，申込者の銀行での取引データや信用スコアを入手し，融資可か不可かを判断し，その結果をローン申込者に通知する.

■ 想定される倫理的な問題

- AIによる融資結果に人種や性別のバイアスがある.

Case2

■ AIシステムの利用シーン

- 銀行のローン審査業務において，従来融資担当者が行っていた融資判断をAIがサポートする.

■ AIシステムの構成

- ローン申込者は，ローン審査AIサービスアプリに，申込者情報，融資金額および返済期間を入力する.
- ローン審査AIは入力された情報を元に，申込者の銀行での取引データや信用スコアを入手して，融資可か不可かを判断し，その結果を融資担当者に通知する.
- 融資担当者は，AIの結果を参考に融資の最終判断をし，ローン申込者に回答する.

■ 想定される倫理的な問題

- AIによる融資結果に人種や性別のバイアスがある.

6. ローン審査AI

ユースケース概要（Case1/Case2共通）

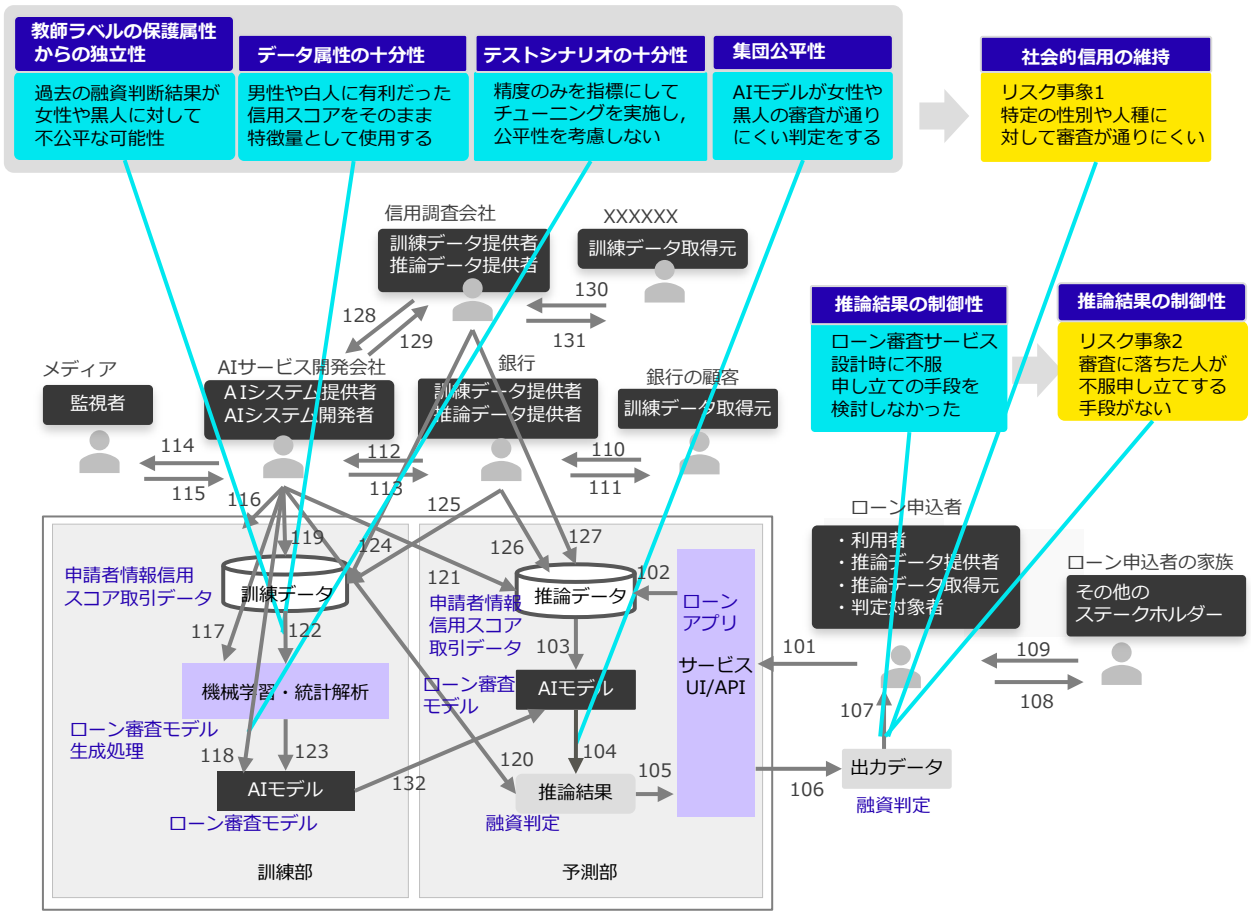
大項目	中項目	内容
業種		金融・保険業
目的		ローン審査にかかる時間の短縮
サービス	サービス概要	短時間のローン審査サービス
	顧客毎のカスタマイズの有無	無し
	要件	無し
利用シーン		ローン審査申込後短時間で審査結果を申込者に回答する
ステークホルダーと役割	AIサービスプロバイダ	銀行
	開発者	AIサービスベンダ
	ビジネス利用者	銀行
	訓練データ提供者	銀行
	訓練データ取得元	銀行
	訓練データ取得時の関係者	銀行の顧客
	推論データ提供者	ローン申込者
	推論データ取得元	ローン申込者
	推論データ取得時の関係者	不明
	消費者的利用者	ローン申込者
	監視者	メディア，金融当局
	サービスUI/API提供者	AIサービスベンダ
	判定対象者	ローン申込者
	サービス認可者	金融当局
	その他のステークホルダー	ローン申込者の家族
Human-in-the-loopの有無		Case1: 無し（AIの判断結果をローン申請者に直接回答） Case2: 有り（AIの判断結果を参考に融資担当者が最終判断し，ローン申請者に回答）
既存手段の有無		有り（融資担当者による判断）

6. ローン審査AI

大項目	中項目	内容
AIタスク	タスク	機械学習や統計解析を用いて，ローン申込者への融資可/不可を判断する
	問題種別	分類問題
	出力	融資可，不可
	技術	自前+OSS
	AIタスクの実施に必要なモデルの内訳	ローン審査モデル
	学習データの更新・追加学習の有無	無し
	リアルタイム性	有り
	AIタスクの検証（既存手段による判断との整合性）	有り
訓練データ	内訳と提供者および取得元	融資内容，取引データ（提供者：銀行，取得元：顧客），信用スコア（提供者：信用調査機関，取得元：不明）
	教師ラベル	返済実績
	保護属性，公平性検証・緩和に使う属性	性別，年齢，人種
	個人情報の有無	無し
	データ取得時の留意事項	不明
	データの保存	不明
推論データ	内訳と提供者および取得元	申込内容，取引データ，信用スコア（提供者と取得元：ローン申込者）
	保護属性，公平性検証・緩和に使う属性	性別，年齢，人種
	個人情報の有無	有り
	データ取得時の留意事項	不明
AIシステムの出力	出力時の留意事項	不明
引用元，参考記事		ISO IEC TR 24030:2021 Artificial Intelligence(AI) - Use cases, https://www.iso.org/standard/77610.html

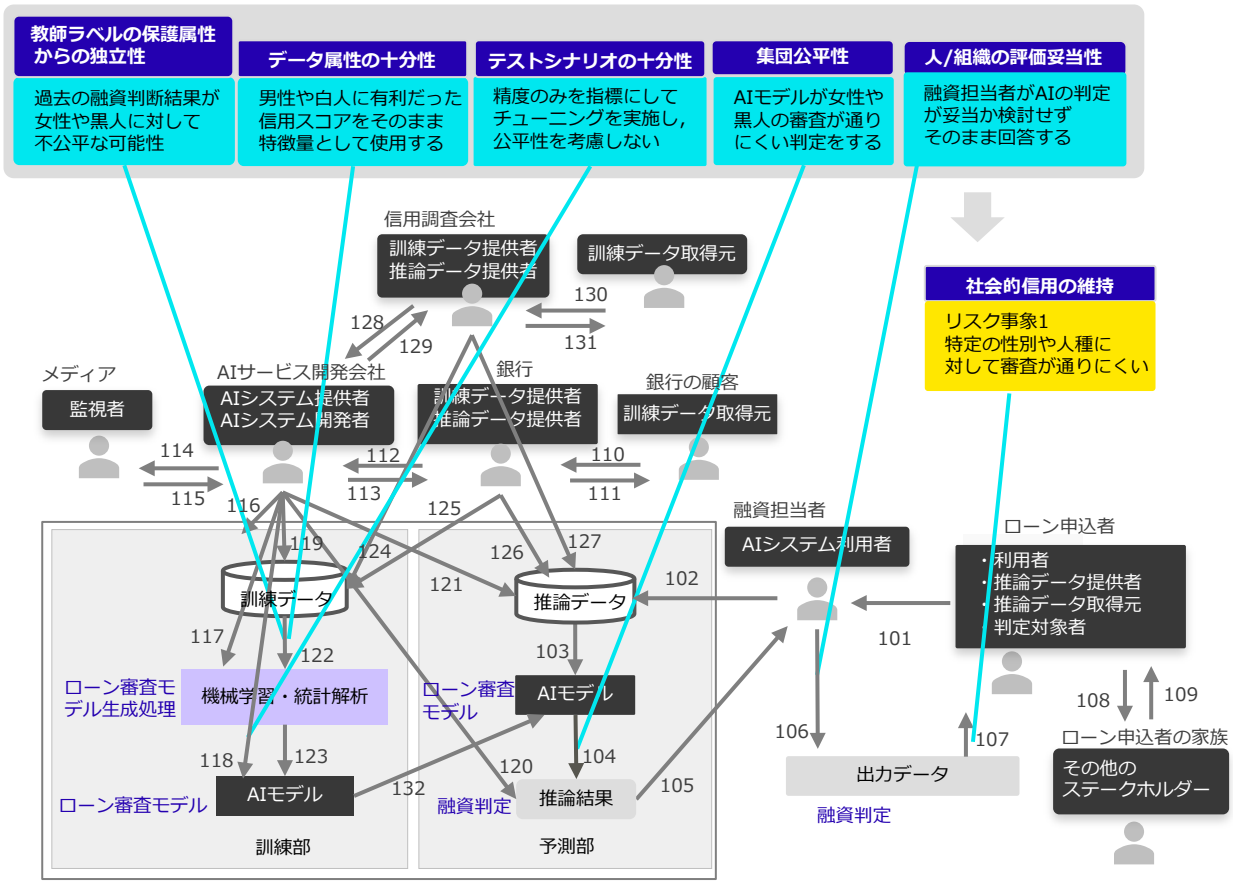
6. ローン審査AI

分析図 Case1



6. ローン審査AI

分析図 Case2



6. ローン審査AI

リスクの整理 Case1

リスク事象	リスク要因	AI倫理特性	インタラクションID
女性や黒人に対して審査が通りにくい		社会的信用の維持	107
	AIモデルが女性や黒人の審査が通りにくい判定をする	集団公平性	104
	精度のみを指標にしてチューニングを実施し，公平性を考慮しない	テストシナリオの十分性	118
	男性や白人に有利だった信用スコアをそのまま特徴量として使用する	データ属性の十分性	122
	過去の融資判断結果が女性や黒人に対して不公平な可能性がある	教師ラベルの保護属性からの独立性	122
審査に落ちた人が不服申し立てする手段がない		推論結果の制御性	107
	ローン審査サービス設計時に不服申し立ての手段を検討していない	推論結果の制御性	107

リスクの整理 Case2

リスク事象	リスク要因	AI倫理特性	インタラクションID
女性や黒人に対して審査が通りにくい		社会的信用の維持	107
	AIモデルが女性や黒人の審査が通りにくい判定をする	集団公平性	104
	精度のみを指標にしてチューニングを実施し，公平性を考慮しない	テストシナリオの十分性	118
	男性や白人に有利だった信用スコアをそのまま特徴量として使用する	データ属性の十分性	122
	過去の融資判断結果が女性や黒人に対して不公平な可能性がある	教師ラベルの保護属性からの独立性	122
	融資担当者がAIの判定が妥当か検討せずそのまま回答する	人/組織の評価妥当性	106

7. 果物等級判定

■ AIシステムの利用シーン

- 果物の画像から等級を判定するAIを用いて，農業団体が農家から集荷した果物を等級判定し，等級に見合った支払いをする

■ AIシステムの構成

- 訓練部：等級判定AIモデルは，果物画像を学習して生成される
- 予測部：農家から集荷された果物は，ベルトコンベアで運ばれ果物1個毎に，画像が撮影され，その等級をAIが判定する．判定結果は後処理において，農家毎に集約され農家毎の支払金が決定される

■ 想定される倫理的な問題

- 農業団体が，農家の合意を得ずに，等級判定結果のデータを分析して農家のランク付けを推察する可能性がある

ユースケース概要

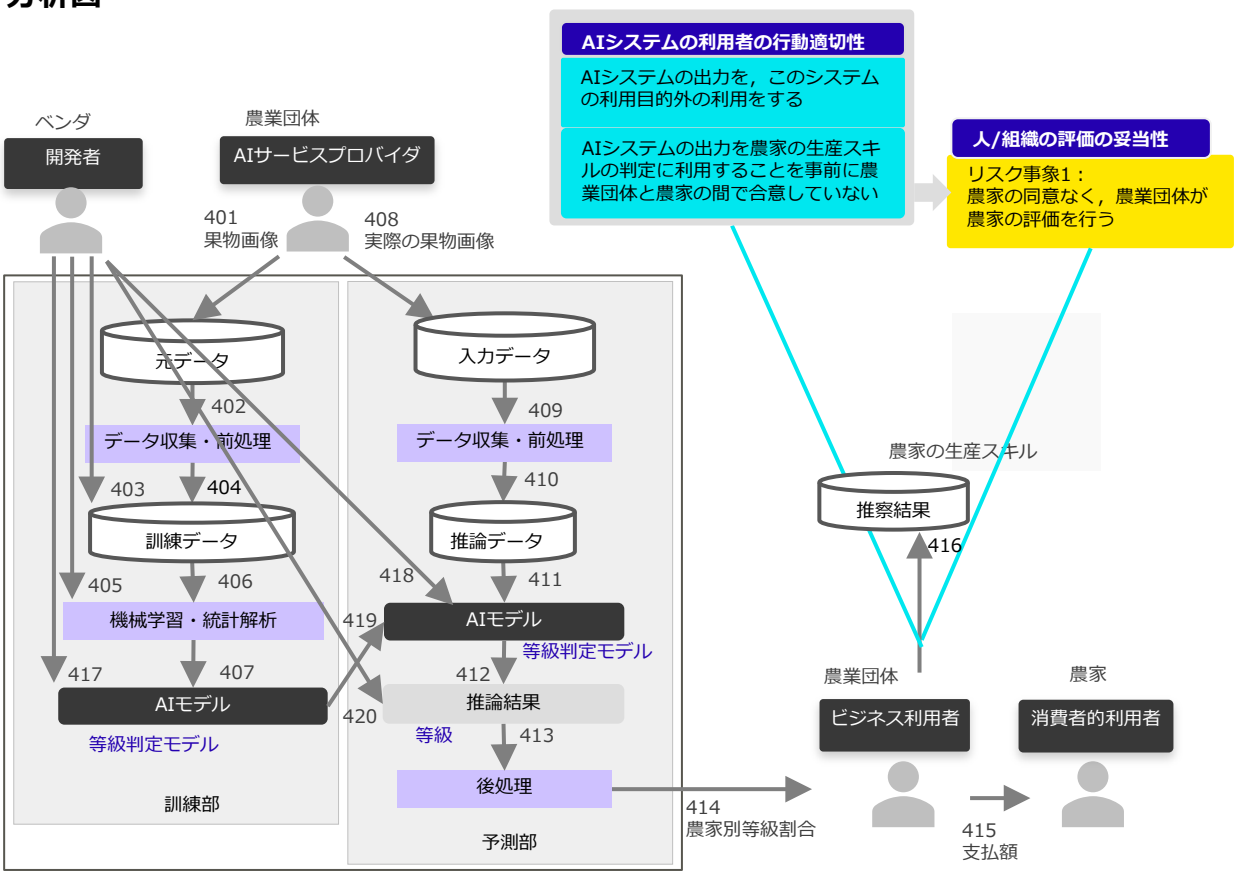
大項目	中項目	内容
業種		農業
目的		農家への支払い額の決定
サービス	サービス概要	農業団体が果物の等級判別AIを用いて，農家毎の支払額を決定する
	顧客毎のカスタマイズの有無	無し
	要件	無し
利用シーン		農家から集荷した果物の等級をAIシステムで判断．農家の等級別割合から農家への支払額を決定.
ステークホルダーと役割	AIサービスプロバイダ	AI開発ベンダ
	開発者	AI開発ベンダ
	ビジネス利用者	農業団体
	訓練データ提供者	農業団体
	訓練データ取得元	農家
	訓練データ取得時の関係者	無し
	推論データ提供者	農業団体
	推論データ取得元	農家
	推論データ取得時の関係者	無し
	消費者的利用者	農家

7. 果物等級判定

大項目	中項目	内容
ステークホルダーと役割	監視者	無し
	サービスUI/API提供者	無し
	判定対象者	農家
	サービス認可者	無し
	その他のステークホルダー	無し
Human-in-the-loopの有無		無し
既存手段の有無		有り
AIタスク	タスク	果物の画像から等級を判定する
	問題種別	分類
	出力	等級
	技術	自前＋OSS
	AIタスクの実施に必要なモデルの内訳	果物等級判定モデル
	学習データの更新・追加学習の有無	無し
	リアルタイム性	有り
	AIタスクの検証（既存手段による判断との整合性）	農業団体のベテラン作業員による等級判別作業
訓練データ	内訳と提供者および取得元	果物画像(提供者と取得元：農業団体)
	教師ラベル	果物等級
	保護属性，公平性検証・緩和に使う属性	無し
	個人情報の有無	果物を出荷した農家
	データ取得時の留意事項	無し
	データの保存時の留意事項	無し
推論データ	内訳と提供者および取得元	果物画像（提供者：農業団体，取得元：農家）
	保護属性，公平性検証・緩和に使う属性	無し
	個人情報の有無	果物を出荷した農家
	データ取得時の留意事項	無し
AIシステムの出力	出力時の留意事項	無し
引用元，参考記事		無し

7. 果物等級判定

分析図



リスクの整理

リスク事象	リスク要因	AI倫理特性	インタラクションID
農家の同意なく，農業団体が農家の評価を行う		人/組織の評価妥当性	416
	AIシステムの出力を，このシステムの利用目的外の利用をする	AIシステムの利用者の行動適切性	416
	AIシステムの出力を農家の生産スキルの判定に利用することを事前に農業団体と農家の間で合意していない	AIシステムの利用者の行動適切性	416

参考文献

[1] https://incidentdatabase.ai/cite/16

[2] https://incidentdatabase.ai/cite/37

[3] https://www.eismd.eu/wpcontent/uploads/2019/11/AI4People%E2%80%99s-Ethical-Framework-for-a-Good-AI-Society_compressed.pdf

[4] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[5] https://oecd.ai/en/ai-principles

[6] https://ethicsinaction.ieee.org/

[7] https://www.soumu.go.jp/main_content/000637097.pdf

[8] https://www8.cao.go.jp/cstp/aigensoku.pdf

[9] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

[10] https://partnershiponai.org/

[11] https://incidentdatabase.ai/

[12] https://www.iso.org/standard/77610.html

[13] https://www.jil.go.jp/kokunai/statistics/databook/2019/03/d2019_T3-01A.pdf