FUJITSU

# AI倫理影響評価

## ～原則から実践へ～

White paper
2022

# エグゼクティブサマリ

社会のあらゆる場面においてAIの適用が広がる一方で、AIが引き起こす倫理的な問題が明るみになってきています。このような背景により、欧州をはじめとする国や組織から、AI倫理原則やAI倫理ガイドラインが策定されています。また、倫理的な問題に対処するための実践的な技術の開発も進んでいます。しかしながら、原則と実践にはまだギャップがあり、原則から実践へつなげていくことが課題です。

我々は、この課題に対して、AI倫理ガイドラインに基づいて、AIが人や社会に対して与える倫理的な影響を評価する方式を開発しました。まず、このギャップを埋める手がかりを得るため、これまでに発生したAIが引き起こした倫理的な問題を精査しました。その結果、問題が、AIとステークホルダーの間、ステークホルダー同士の間といったインタラクションで生じていることがわかりました。AI倫理影響評価では、この点に着目し、倫理ガイドラインで示されている信頼できるAIの要件を、AIのシステムに現れるインタラクションに対応づけて、AIによる倫理的な問題を網羅的に洗い出します。この影響評価を設計・監査段階で実施することで、AIが深刻な社会的な問題を引き起こさないように事前に回避する対策を講じることができるようになります。

このAI倫理影響評価の実施手順を示した実践ガイド、および、代表的な事例への適用例を無償公開します。これを第一歩として、同じ思いを共有する国や組織、また、技術のみならず法律や哲学など、多様な知見や視点を有する方々と共に、信頼できるAIの提供を目指します。

# 1. イントロダクション

## AIの課題

金融取引、医療や雇用など社会のあらゆる場面において広くAIが適用されています。担当者が課税申告などの書類をひとつひとつ修正していた作業を瞬時に自動で行うなど、作業負荷を軽減したり、膨大なデータに基づいて患者のがん種別に応じた適切な治療薬を選択するなど、これまで容易にできなかった判断を可能にしたりしています。一方で、AIが引き起こす倫理的な問題が明るみに出て、大きく報道されるようになっています。顔認識AIが人種差別的な結果を出したり、人材採用AIが性差別的な結果を出すために運用停止となったりしています。このような問題が発生すると、AI を提供した企業や組織が社会的な信用を失うだけでなく、AI の利用者や社会に対する負の影響も大きくなります。

富士通は、「人をすべての中心に置く」という考えを長年にわたって大事にしてきました。この考えに基づいて、社会課題の解決やビジネスの変革を進めています。AIの社会実装においても、富士通はこの人間中心のアプローチで、安心してAIが利用できること、すなわち信頼されるAIを実現することを目指しています。

## 信頼できるAIに向けた取組み

なぜ、AIは問題を起こすのでしょうか。機械学習は過去のデータに基づいて行われることから、これまで社会に潜んでいた差別や不公平を含む偏ったデータを学習することで、AIによる判断にも偏りが生じることがあります。また、どのような根拠で導いたのかが明らかではないAIによる判断を、人が介在することなくそのまま利用することへの不安が生じています。

このような背景から、欧州をはじめとする国や組織は、信頼できるAIを普及させるために、基盤となる考え方を示すAI倫理原則（以下、倫理原則）や、その倫理原則を適用するための要件を示すAI倫理ガイドライン（以下、倫理ガイドライン）の策定を行っています。さらに、いくつかの特定のAIの使用を禁じたり、法執行や公共サービスをはじめとする、人を対象とした重大な判断に関わるシーンでのAIの使用に厳しい制限を加えたりする法制化の動きも始まっています。一方で、AIによる判断の偏りを緩和・是正する技術や、AIの結果を導出するに至った根拠を説明する技術などの開発も進んでいます。

このように、倫理ガイドラインの策定を通じて信頼できるAIが満たすべき要件を明らかにすることと、起こり得る問題に対処できるようにすることの両面から、信頼できるAIを提供するための取組みが進められています。

## 「実践」に向けた我々の取組み：AI倫理影響評価

倫理原則と、これに則るための倫理ガイドラインは、信頼できるAIの理念や、それが満たすべき要件を明確にする上で重要です。同様に、AIがこれらを遵守するように制御する実践的な技術も不可欠です。しかしながら、原則と実践には大きなギャップがあり，原則と実践をつなぐためにはどうすればよいかという問題に対する答えは必ずしも明確ではありません。

そこで、我々は原則と実践のギャップを埋める手がかりとして、AIとこれを取り巻くステークホルダーたちの間のインタラクションに注目することにしました。なぜなら、公平性を含む倫理的な問題が社会的な関係性において発生しているように、AIにまつわるインシデントは、 AIとステークホルダーの間、ステークホルダー同士の間、あるいはその連鎖において観察されるためです。

我々が提供するAI倫理影響評価は、倫理ガイドラインを分析し、過去のインシデントと照らし合わせることで体系化した知識に基づいて、あなたが提供するAIが、深刻な社会的な問題を引き起こさないように設計・監査するための情報を提供するものです。具体的には、AIとこれを取り巻く人々のインタラクションを明確化したのちに、提供する手順に従って分析することで、倫理ガイドラインを遵守し、これまでに知られているインシデントを回避する情報を得ることを可能にします。

## AI倫理影響評価で目指したいこと

我々は、Partnership on AI※1（以下、PAI）が提供するインシデントDBに登録されている事例に対して、AI倫理影響評価を実施しました。PAIは、AIによる倫理的な問題に取り組む非営利組織であり、そのインシデントDBは、AIの関係者が失敗から学ぶ機会を提供しています。そのDB内の事例に対して、欧州のAIハイレベル専門グループの提供する倫理ガイドライン「Ethics Guidelines for Trustworthy AI※2（以下、Trustworthy AI）」の要件との不整合を検知してみた結果、インシデントがインタラクションに対応づけられ、これらを検出できることが確かめられました。

その成果物である、Trustworthy AIを具体化してインタラクションと対応づけて整理したAI倫理モデルと、それを使ってAI倫理影響評価を実施する手順書で構成されるAI倫理影響評価実践ガイド，および，事例への適用結果で構成されるAI倫理影響評価適用例を無償公開します。様々なAI事例にこの影響評価を適用することで、AIによる倫理的な問題が起きることを事前に防ぎ、安心してAIを提供できるようにすることを目指します。

※1）https://partnershiponai.org/
※2）https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

# 2. AI倫理に関する動向

## AI倫理原則・AI倫理ガイドライン策定の動向

AIが引き起こす倫理的な問題への意識の高まりをうけて、信頼できるAIを普及させるための基本理念を示した倫理原則や倫理ガイドラインの策定が行われています。日本では、内閣府より「人間中心のAI社会原則※3 」が公開され、この倫理原則を含んだ、AIに関する初の国際的な政策ガイドラインである「OECD　AI原則※4」が公開されました。このOECD原則をベースとして、より実践に近いガイドラインとして、総務省「AI利活用ガイドライン※5」や、経産省「AI原則実践のためのガバナンスガイドライン※6」が発行されています。

欧州では、信頼できるAIの開発を促すためのガイドラインとして、AIハイレベル専門グループより「Ethics Guidelines for Trustworthy AI」が公開されています。米国では、IEEEから倫理的なAI開発を促進するための指針として「Ethically Aligned Design※7」が公開されています。

国や組織の価値観に応じて倫理ガイドラインが定められるため、それぞれの特色が表れてはいますが、原則で示されている理念について共通点が認められます。

## AIに対する法規制の動向

AIによる社会への混乱を避けるため、法規制の動きも進んでいます。欧州では、欧州委員会からAI規則案※8が公表されました。このAI規則案では、人の潜在意識への操作、社会的スコアリングの利用、公的空間での法執行目的の遠隔生体認証などを「受容できないAI」と位置づけて禁止しています。また、人の生体認証や分類、重要インフラへの適用などを「ハイリスクAI」としてリストアップし、それらの利用には多くの義務を課し、それを違反した場合には膨大な罰金を課すものとなっています。米国では、「Facial Recognition and Biometric Technology Moratorium Act※9」と呼ばれる連邦政府関係者が顔認識技術を使用することを禁止する法案が提出されました。その他、サンフランシスコ市が警察の顔認識技術の使用を禁止するなどの動きがあります。

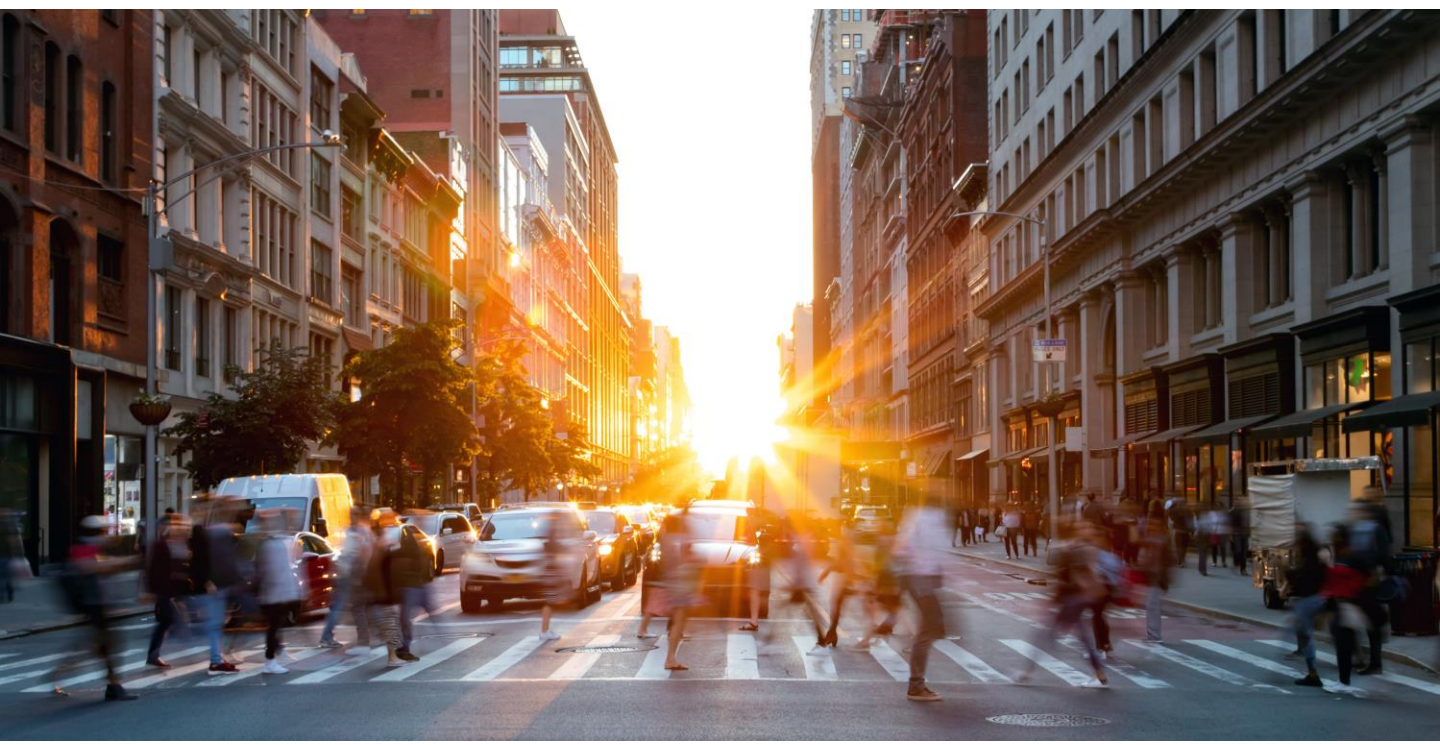
## AI開発ガイドライン策定の動向

国内外で示されている倫理ガイドラインをベースに、より開発現場に即した開発ガイドラインや開発プロセスを策定し、AIの開発段階から倫理的な問題を起こさないような体制の構築が進められています。国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託事業として国立研究開発法人産業技術総合研究所が中心となって進めている「人と共に進化する次世代人工知能に関する技術開発事業」においては、AIの品質にフォーカスした「機械学習品質マネジメントガイドライン※10」が公開されています。これはAI品質についての開発ガイドラインですが、第2版では公平性についての記載が強化されています。 また、AI開発ベンダーとなる各社でも独自に開発ガイドラインの作成を進めています。

## AI倫理の問題に対処する技術開発の動向

倫理ガイドラインに遵守するための技術として、AIの公平性に関わる問題に対処するための技術開発が進められています。公平性については、いくつかの基準が定義されています。例えば、公平として扱いたい属性（例えば、男性と女性）について、AIの出力（例えば、人材採用AIであれば採用・不採用）の確率を均等にするものや、正解データにおいてAIの出力が採用となる確率を均等にするものがあります。これらの様々な基準に従うような機械学習アルゴリズムが盛んに研究されています。

また、AIがどのように結果を導き出したか説明する「説明可能AI」の技術開発も進められています。説明可能AIでは、例えば、画像認識においてAIが画像のどの領域に注目したかを示したり、推論対象のどんな属性がAIの判断に寄与したのかを定量的に示したりするものが開発されています。



## 望まれる今後の動き

このように、AI産業を政策の中心に据える国や地域において、学問領域や産官学を跨いだAI倫理のあり方と実現に向けた多くの成果が知られるようになりました。これらは、必ずしも当該国や地域にとどまらず、世界的に、様々な産業に影響することになるでしょう。社会的に大きなインパクトをもたらすAIが、人間中心の価値観に沿った産業であり続けるために、今後もそれぞれの専門領域における最先端の研究や、社会との間での建設的なコミュニケーションが強く望まれます。それと同時に、専門領域の知見を相互に接続していくために、新しい視点が望まれると我々は考えています。

※3）https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/aigensoku.pdf
※4）https://www.oecd.org/tokyo/newsroom/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence-japanese-version.htm
※5）https://www.soumu.go.jp/main_content/000637097.pdf
※6）https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_6.pdf
※7）https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
※8）https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682
※9）https://www.congress.gov/bill/117th-congress/house-bill/3907
※10）https://www.digiarc.aist.go.jp/publication/aiqm/

# 3. AI倫理影響評価

## AI倫理影響評価のアイデア

この数年間にわたるAI倫理に関する哲学、技術、および政策のレベルでの、AIは倫理的にどうあるべきかという根本的な議論に続き、AI倫理を「原則から実践へ」と推し進めていくことが、議論の対象として浮上してきました。原則を実践したAIを実現するためには、AIのひとつひとつを信頼できるものに育てていくか、倫理的な側面を取り扱うことができるよう技術を強化するか、少なくとも2通りのアプローチがあり、両者を組み合わせた現実的なアプローチを採用することが考えられます。

我々は、これらを補完するアプローチとして、AIが倫理的に正しく振舞う設計と、そのように機能することの監査の実現のために、様々な事例に共通する視点を検討しました。過去のインシデントを観察すると、AIとこれを取り巻くステークホルダーたちの間のインタラクションおよびその連鎖において発生していることがみえてきました。この観察結果に基づけば、AIの構成要素やステークホルダーのインタラクションを中心に調べることで、AIが倫理ガイドラインに遵守していることの客観的な評価の可能性が示唆されます。我々は、ソフトウェア要求工学に立脚し、倫理ガイドラインと、AIの具体事例の詳細な分析を通じて、AI倫理影響評価を開発しました。以下では、このAI倫理影響評価について概説します。

## AI倫理影響評価とは

AI倫理影響評価は、次のようなプロセスで、倫理ガイドラインに基づいて、AIによる倫理的な問題とそれを引き起こす要因をリスクとして洗い出します。

まず、倫理ガイドラインに基づく倫理要件を、AIのシステム上のどこで確認すればよいかを体系的に整理したAI倫理モデルを作成します（**図1**）。具体的には、ソフトウェア工学の要件定義の手法を適用して、倫理ガイドラインの概念的な記述を具体的な倫理要件に整理します。これに、どこのインタラクションで問題が発生しているかという過去のインシデントの分析の結果を紐づけます。このAI倫理モデルは、一つの倫理ガイドラインに対して一度だけ作成すればよく、様々なAIの影響評価において利用することができます。

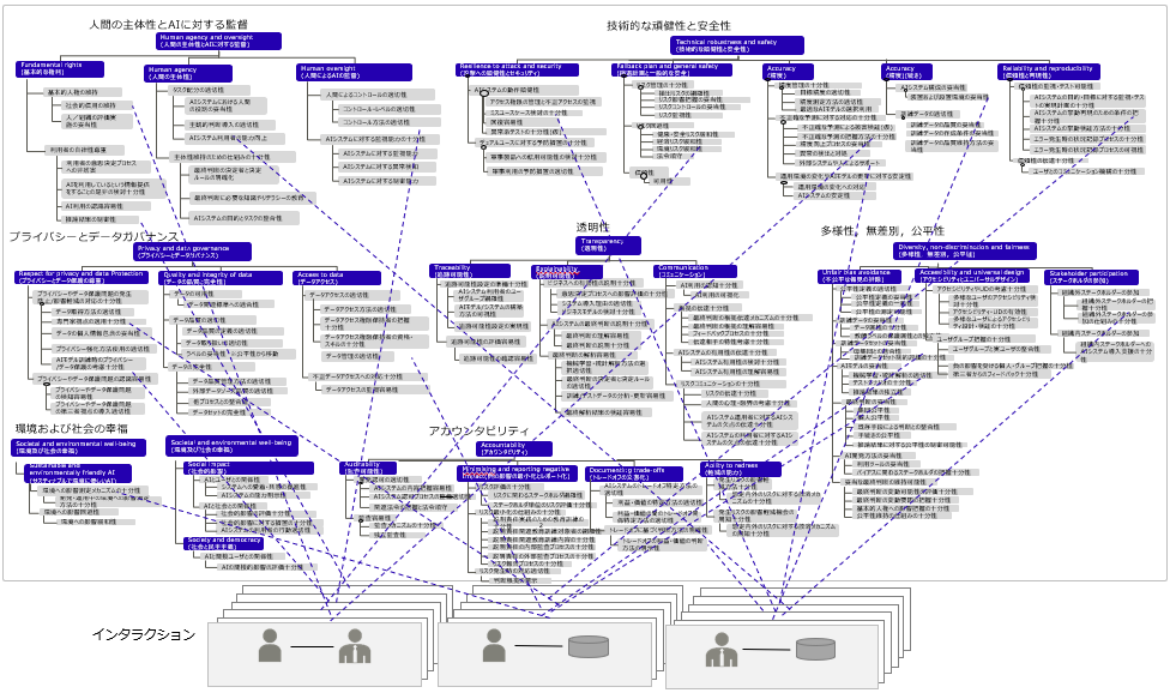



**図1：AI倫理モデル**

次に、AIのシステム上のインタラクションと、AI倫理モデルを対応づけることで、検討すべき倫理要件から起こり得るリスクを抽出します（図２）。AIのシステムは、ソフトウェア開発の設計時に用いられるモデリング手法に基づく、AIのシステムの構成要素と関係するステークホルダーを配置したシステム図で表します。このシステム図は、AI事例から抽出された数パターンのシステム図のバリエーションとして作成することができます。このことは、スタンフォード大学mediaXとの共同ワークショップ※11において確認されました。システム図ができれば、その図に現れるインタラクションに応じて、AI倫理モデルから具体的な倫理要件を抽出することで、系統的にリスクを洗い出します。

このAI倫理影響評価を実施するために、我々は、システム図を作成するときに利用するシステム図のパターンシートを付録として載せた実施手順書と、AI倫理モデルで構成される、AI倫理影響評価実践ガイドを無償公開します。

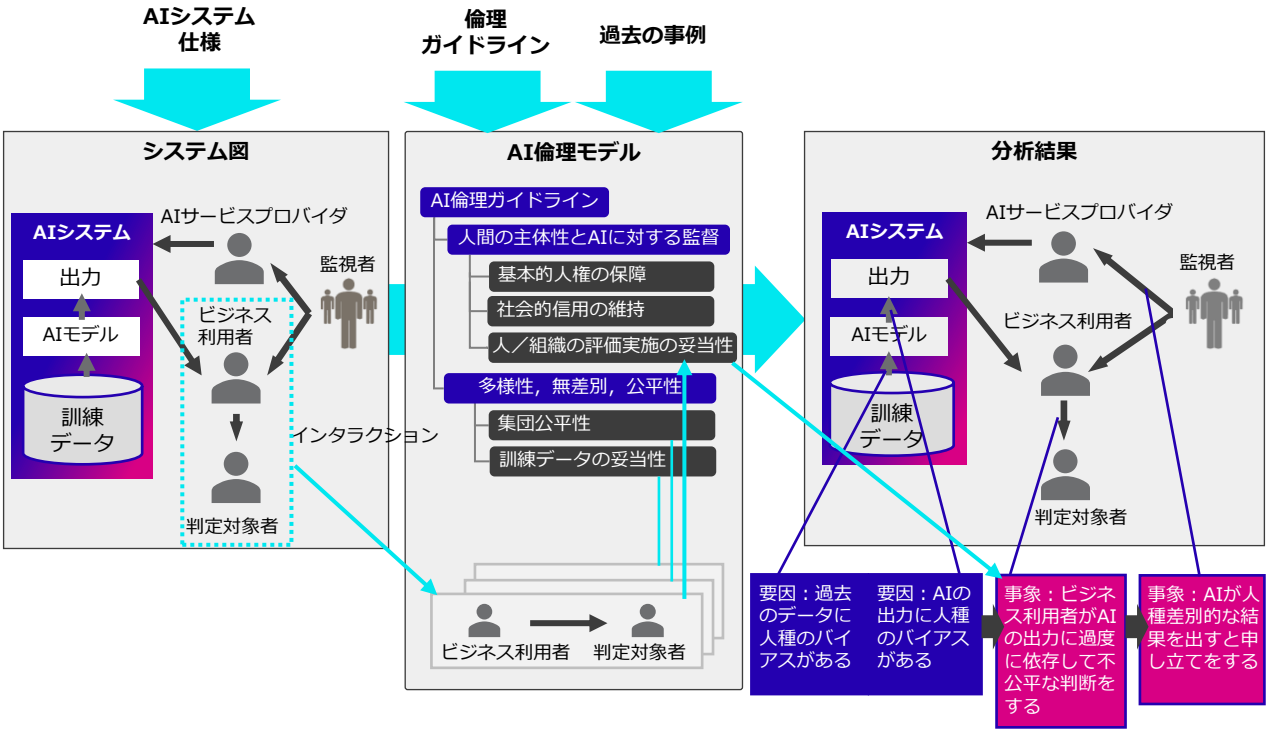



**図2：AI倫理影響評価の全体像**

※11）https://mediax.stanford.edu/event/ai-ethics-workshop/

## AI倫理影響評価の効果

AI倫理影響評価を実施することによってわかることを、架空の再犯リスク予測AI事例に適用した分析結果（**図3**）で説明します。再犯リスク予測AIは、過去の被告人のデータを学習し、新たな被告人の情報から、その被告人の再犯リスクを予測します。裁判官は、このAIの予測結果を、被告人の仮釈放の有無や量刑の判断を下す際に参考にします。**図3**に示す抽出されたリスクから、過去のデータの属性によるバイアスや不公平な予測結果を要因として、裁判官の判断が偏る、さらにはそれがメディアに取り上げられる、という可能性を想定できます。米国の一部ではこれに近い再犯リスク予測AIが実際に採用されており、メディアが「再犯リスク予測AIが差別的である」という記事を公開し[注1]、議論を巻き起こすというインシデントがありました。米国におけるインシデントについては、記事による情報しかないため、架空の事例とは少し異なるかもしれませんが、この架空の事例の分析結果からは、インシデントを引き起こしたリスクを、AI倫理影響評価で抽出できていることがわかります[注2]。また、リスクが起きているインタラクションも可視化されるため、具体的な施策を立てやすくなります。

このように、AI倫理影響評価は手順書に従って系統的に行うため、倫理ガイドラインの範囲において、もれなく影響を評価することができます。

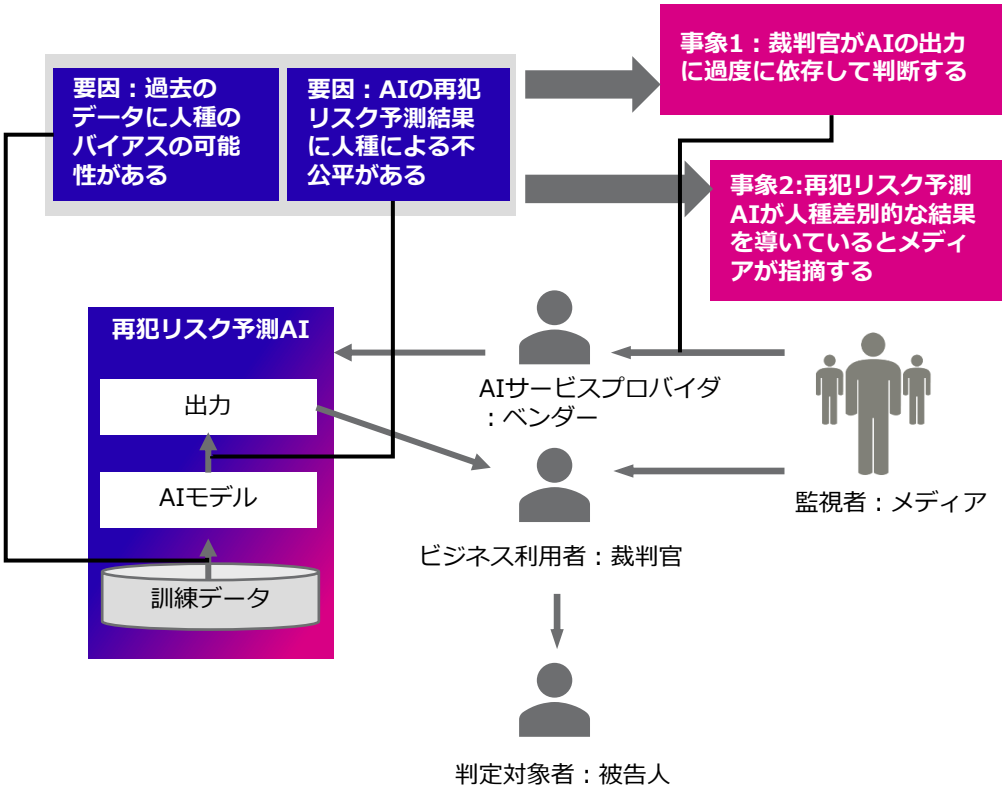



**図3：再犯リスク予測AIのAI倫理影響評価適用例**

# 3. AI倫理影響評価

AI倫理影響評価を未知のインシデントに適用して、そのAI倫理影響評価の効果を検証しました。PAIが提供するAIに関するインシデントDBに登録されている150件以上のAIインシデント事例を、業種およびAIアプリケーション種別で分類した中から、未知のインシデントとして約15事例を選択しました。これらに対してAI倫理影響評価を実施した結果、インシデントとそれを引き起こす要因となるリスクが、AIとステークホルダーとのインタラクションに対応づけられ、それらをすべて抽出できることを確認しました。なお、この検証で使用したAI倫理モデルはTrustworthy AIから作成したものです。我々は、このAI倫理モデルと分析手順書で構成されるAI倫理影響評価実践ガイドと、代表的な事例への適用例で構成されるAI倫理影響評価適用例を無償公開します。

# 4.結び

AI倫理影響評価は、AIの倫理的な影響を系統的かつ網羅的に評価し、AIによる倫理的な問題がどこで起こり得るのか、という情報を提供します。これを，AIの開発あるいは提供前に適用することで、AIが引き起こす倫理的な問題への事前の対処を講じることができるようになります。

我々は、AI倫理影響評価を発展させ、信頼できるAIに対する共通の理念を有する国や組織と、AIに関する技術だけではなく、法律や哲学などの多様な知見や視点を取り入れて、信頼できるAIの社会への普及を目指します。

## 注記

[1] Machine Bias - There’s software used across the country to predict future criminals. And it’s biased against blacks, by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica, May 23, 2016 https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] ここで紹介した事例および分析図は、一部、推測による記載も含みます。通常、富士通のAI倫理影響評価では、事例の内容を明らかにして分析を進めていきますが、今回は事例として詳細が公表されていない部分があるため、必ずしも事実が本稿の記載通りとは限らない点、ご了承ください。

**富士通株式会社**
研究本部 AI倫理研究センター
E-mail： fj-labs-aie-info@dl.jp.fujitsu.com

**商標について**
記載されている製品名などの固有名詞は、各社の商標または登録商標です。

**免責事項**
富士通は、本方式により洗い出されたもの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、本方式を利用する各団体・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。



2022年2月発行　V1.0