

欧州データ保護監督官
EU の独立したデータ保護機関

人工知能システムのリスクマネジメントに関するガイダンス

2025 年 11 月 11 日

目次

エグゼクティブサマリー	3
1 序論	4
1.1 目的	4
1.2 適用範囲	4
1.3 対象読者	5
2 リスクマネジメント手法	6
3 AI ライフサイクルの理解	8
3.1 AI システムの定義	8
3.2 AI システムのライフサイクル.....	8
3.3 AI システムの調達	9
4 解釈可能性と説明可能性は不可欠な条件.....	11
4.1.1 リスク 1： 解釈不能または説明不能な AI システム	12
5 主要なデータ保護原則に関連するリスク.....	14
5.1 公正性の原則	14
5.1.1 リスク 1： 訓練用個人データの品質不足によるバイアス.....	15
5.1.2 リスク 2： トレーニング個人データにおけるバイアス.....	17
5.1.3 リスク 3： トレーニング個人データへの過学習	18
5.1.4 リスク 4： アルゴリズムバイアス	19
5.1.5 リスク 5： 解釈バイアス.....	21
5.2 正確性の原則	22
5.2.1 EUDPR における正確性の法的意味	22
5.2.2 AI 開発における正確性の統計的意味	22
5.2.3 リスク 1： 不正確な個人データの出力.....	22
5.2.4 具体例： データドリフトと入力個人データの品質劣化による不正確な出力	24
5.2.5 リスク 2: AI システムプロバイダからの不明確な情報	25
5.3 データ最小化の原則.....	26
5.3.1 リスク 1： 個人データの無差別な収集と保存	27
5.4 セキュリティの原則	28
5.4.1 リスク 1： AI システムの出力によるトレーニング個人データの開示	28
5.4.2 リスク 2： 個人データの保存と個人データ漏えい.....	30
5.4.3 リスク 3： アプリケーションプログラミングインターフェース（API）を通じた個人データ漏洩	31
5.5 データ対象者の権利.....	32
5.5.1 リスク 1： 処理される個人データの不完全な特定	32
5.5.2 リスク 2： 不完全な訂正または消去	33
6 結論	35
附属書 1： 評価指標	36
附属書 2： 懸念事項とリスクの概要	42
附属書 3： AI ライフサイクル開発の各段階におけるチェックリスト.....	43
AI システムの開発.....	43
AI システムの調達.....	45

エグゼクティブサマリー

欧州連合の機構、団体、事務所及び機関（EUI）によるパーソナルデータの処理を伴う AI システムの開発、調達及び展開は、プライバシー及びデータ保護を含むがこれらに限定されない、データ対象者の基本的権利及び自由に対する重大なリスクをもたらす。規則 2018/1725（EUDPR）の基盤である¹において、第 4 条(2)（行政上の個人データ）および第 71 条(4)（業務上の個人データ）に規定される説明責任の原則は、EUI に対し、これらのリスクを識別・緩和するとともに、その方法を実証することを要求している。これは、複雑なサプライチェーンの産物である AI システムにおいて特に重要である。こうしたシステムでは、異なる立場で個人データを処理する複数の主体が関与することが多いからだ。

本ガイダンスは、データ管理者として行動する EUI が、こうしたリスクの一部を識別・緩和するための指針となることを目的とする。具体的には、EUDPR で規定される特定のデータ保護原則（公平性、正確性、データ最小化、セキュリティ、データ対象者の権利）への非準拠リスクに焦点を当て、管理者が実施すべき緩和策は技術的性質を持つ場合がある。したがって、本ガイダンスに記載された技術的対策は決して網羅的ではなく、EUI が自らの特定の処理活動によって生じるリスクを独自に評価する義務を免除するものではない。また、その際、リスク発生の可能性や深刻度の順位付けは行わない。

まず、本文書は ISO 31000:2018 に基づくリスクマネジメント手法の概要を示す（第 2 節）。次に、AI システムの典型的な開発ライフサイクルと調達プロセスにおける各段階を概説する（第 3 節）。第三に、解釈可能性と説明可能性という概念を、本ガイダンスで扱う全規定への準拠を左右する横断的課題として考察する（第 4 節）。最後に、前述の 4 つの一般原則（公平性、正確性、データ最小化、セキュリティ）を具体的なリスクに分解し、各リスクの説明と併せて、管理者がリスク緩和のために実施可能な技術的措置を提示する（第 5 節）。

欧州データ保護監督官（EDPS）は、本ガイダンスを AI 法に基づく市場監視当局としての役割ではなく、データ保護監督当局としての立場で発行する。本ガイダンスは人工知能法に影響を与えない。

¹欧州議会及び理事会規則（EU）2018/1725（2018 年 10 月 23 日）欧州連合の機構、団体、事務所及び機関によるパーソナルデータの処理及び当該データの自由な移動に関する自然人の防御、並びに規則（EC）第 45/2001 号及び決定第 1247/2002/EC 号の廃止について[2018] OJ L295/39 <https://data.europa.eu/eli/reg/2018/1725/oj>

1 序論

1.1 目的

本文書は、EU 機構・団体・事務所・機関（EUI）が、人工知能システムの開発・調達・展開時にパーソナルデータの処理によって生じるデータ対象者の基本的権利に対するリスクを識別し緩和するための指針を目的とする。EU 規則 2018/1725（EUDPR）第 3 条(8)項の定義に基づくデータ管理者として行動する EUI を対象とする。²本ガイドラインは、データ保護影響アセスメント及び事前協議に関する「現場での説明責任ツールキット」第 II 部を補完するものである。³

また、2024 年 6 月に発行された「生成的 AI システム利用時のデータ保護コンプライアンス確保に向けた EU 機関による生成的 AI 利用に関する EDPS 指針」を補完するものである。同指針は、EU 機関が生成的 AI システムを開発または利用する際に EUDPR への準拠を確保する方法について実践的な助言を提供している。⁴本文書は、あらゆる種類の AI システムを対象とする点でより広範であると同時に、法的緩和ではなく技術的緩和に焦点を当てる点でより限定的である（第 1.2 節参照）。

本文書は、AI システムで生じうるリスクを識別・対処するための分析的枠組みを提供する。これは影響を受けうるデータ保護原則に沿って構成されている。本枠組みはコンプライアンスガイドラインを構成するものではなく、またそれに依拠すべきではない。本文書の唯一の目的は、データ保護の観点からリスクを体系的に評価することを促進することである。言い換えれば、これは管理者（データ管理者）が実施すべき各 AI システムに対する必要なコンプライアンス評価に代わるものではない。管理者は、本枠組みを活用して識別されたリスクが、EUDPR（EU データ保護規則）に基づく全ての義務を満たすために必要に応じて管理されていることを確保しなければならない。

欧州データ保護監督官（EDPS）は、本ガイダンスをデータ保護監督機関としての役割において発行するものであり、AI 法に基づく新たな市場監視機関としての役割に基づくものではない。

1.2 適用範囲

本ガイダンスの目的上、ISO 31000:2018（⁵）で使用される用語に基づき、「リスク」の概念は「リスク源」、「事象」、「結果」、「制御」という用語で表現される。ここで「リスク源」とは、AI システムの調達、開発、または導入の文脈における個人データの処理を指す。「事象」、「結果」、「管理」という用語で表現される。ここで「リスク源」とは、AI システムの調達、開発、または展開の文脈におけるパーソナルデータの処理を指し、「事象」とは、その処理がデータ対象者の基本的権利と自由を侵害する状況を指し、「結果」とは、これがデータ対象者に与える可能性のある物質的または非物質的損害を指す。⁶

²欧州議会及び理事会規則（EU）2018/1725（2018 年 10 月 23 日）欧州連合の機構、団体、事務所及び機関によるパーソナルデータの処理及び当該データの自由な移動に関する自然人の保護、並びに規則（EC）No 45/2001 及び決定 No 1247/2002/EC の廃止について[2018] OJ L295/39 <https://data.europa.eu/eli/reg/2018/1725/oj>

³EDPS、現場における説明責任 第 II 部：データ保護影響アセスメント及び事前協議、2018 年 2 月、https://www.edps.europa.eu/sites/default/files/publication/18-02-06_accountability_on_the_ground_part_2_en.pdf

⁴欧州データ保護監察機関、生成的 AI と EUDPR。生成的 AI システム利用時のデータ保護コンプライアンス確保に向けた指針（バージョン 2）、2025 年 10 月 28 日、https://www.edps.europa.eu/system/files/2025-10/2510_28_revised_genai_orientations_en.pdf

⁵国際標準化機構、ISO 31000:2018 リスクマネジメント — 指針、第 2 版、2018 年、<https://www.iso.org/standard/65694.html>。

⁶EUDPR 第 46 項は次のように規定する。「自然人の権利及び自由に対するリスクは、個人データの処理によって生じ得る。具体的には、身体的、物質的又は非物質的損害をもたらす可能性のある処理、特に以下の場合である：処理が差別、身元盗用又は詐欺、金銭的損失、名誉

「管理」とは、データ管理者が当該リスクの発生確率を低減し、かつ／または発生した場合のデータ対象者への影響を軽減するために実施可能な緩和策を指す。⁷

EUDPR 第 1 条(2)に基づき、本規則の目的は、パーソナルデータの処理におけるプライバシー及びデータ保護を含むがこれに限定されない、自然人の権利及び自由を防御することである。第 4 条(2)、第 26 条(1)、第 27 条(1)によれば、EU 域内のデータ管理者は、自らの処理活動によって生じるこれらの権利・自由に対するリスクを識別・緩和する責任を負い、その方法を証明しなければならない。これは特に、AI システムの調達・開発・展開において重要である。AI システムのあらゆる悪影響はまだ評価されていないからだ。したがって、データ管理者は各処理活動において、全てのデータ対象者の基本的権利に対するリスクを適切に識別・緩和することが極めて重要である。EUDPR に明示的に定められた規定への準拠は、この目的達成の指標となる。本ガイダンスは、リスクを「規定違反の事象」として概念化することを堅持する。

より具体的には、本ガイダンスは、データ管理者が実施すべき「管理措置」が技術的性質を帯び得る、**厳選されたデータ保護原則**（すなわち公正性、正確性、データ最小化、セキュリティ、および特定のデータ対象者の権利）への非遵守リスクに焦点を当てる。欧州データ保護監督官（EDPS）は、本ガイダンスで概説するリスクと対策のリストが網羅的ではなく、AI システムの調達、開発、展開時に管理者が対処すべき最も差し迫った課題の一部を反映しているに過ぎないことを強調する。

1.3 対象読者

本文書の対象読者は、AI システムの調達、開発、展開に携わる EUI 職員である。これにはソフトウェア開発者、データ 科学者、IT エンジニア、IT プロジェクトマネージャー、データ保護責任者（DPO）、データ保護コーディネーターが含まれる。

毀損、職業上の秘密によって保護される個人データの機密性の喪失、擬似匿名化の不正な解除、その他の重大な経済的・社会的不利益が生じる場合。データ対象者が権利・自由を剥奪される、または個人データに対する管理権限の行使を妨げられる場合。人種・民族的出身、政治的意見、宗教・哲学的信条、労働組合への加盟を明らかにする個人データ、ならびに遺伝データ、健康に関するデータ、性生活に関するデータ、刑事上の有罪判決・犯罪歴または関連する保安措置に関するデータが処理される場合。個人の側面が評価される場合、特に職務遂行能力、経済状況、健康状態、個人の嗜好・関心、信頼性・行動、位置情報・移動経路に関する側面を分析または予測し、個人プロフィールを作成または利用する場合。脆弱な自然人、特に児童の個人データが処理される場合。または処理が大量の個人データを扱い、多数のデータ対象者に影響を及ぼす場合。

⁷ISO 31000:2018（3.8 項）は「リスクを維持および／または変更する措置」と定義される「制御」という用語を使用している。EUDPR は「措置」という用語を使用している。本稿の残りの部分では「措置」という用語を使用する。

2 リスクマネジメント手法

ISO 31000:2018 によれば、**リスクマネジメント**とは組織がリスクを制御するプロセスである（図 1 参照）。この活動の核心は**リスクアセスメント**部分であり、組織はここでリスクを順次識別、分析、評価する。⁸

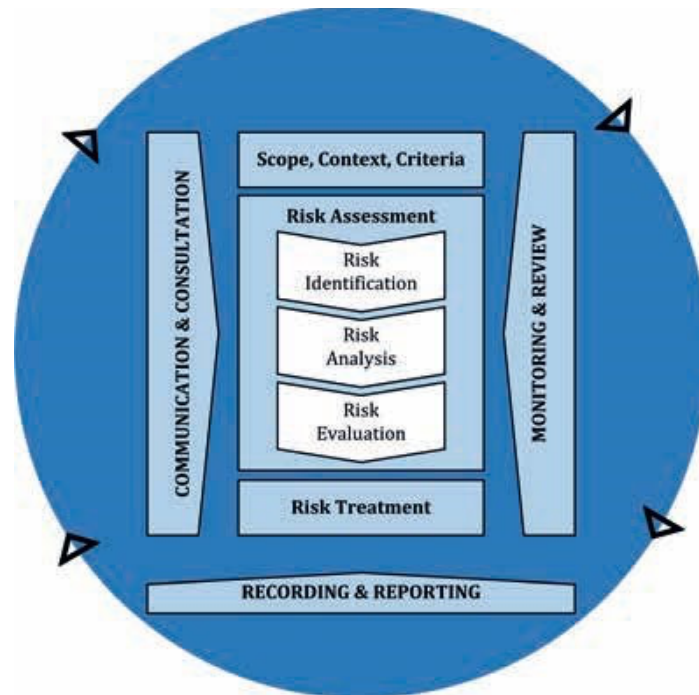


図 1: リスクアセスメント⁹

リスクの識別とは、組織の目標に影響を及ぼす可能性のあるリスクを体系的に認識するプロセスである。この段階では、リスクの発生源、影響範囲、不確実性を招く可能性のある事象や状況を識別することに焦点を当てる。目的は、後続の段階でさらに分析される包括的なリスク登録簿を作成することである。第 1.2 節で既に示唆した通り、本ガイダンスは EUI が追求する目的は、自らがデータ管理者として関与するパーソナルデータの処理がデータ対象者の基本的権利を侵害しないことを確保することであると想定している。

リスク分析は次の段階であり、組織は識別されたリスクを調査し、その性質、発生源、発生可能性、および潜在的な結果を理解する。この段階では、各リスクが発生する可能性と、発生した場合のデータ対象者への影響を判断することを目的とする。定性的な**リスク分析**では、発生可能性と影響のレベルを「非常に低い」、「低い」、「中程度」、「高い」、「非常に高い」の 5 段階で評価できる。¹⁰¹¹これらの要素を定義した後、リスクは発生確率と影響の深刻度の積として評価される（リスク = 発生確率 × 影響）。これは通常、リスクマトリックスで表される（図 2 参照）。

⁸Isabel Barberá, *AI Possible リスク & 緩和 - Named Entity Recognition*, September 2023, https://www.edpb.europa.eu/system/files/2024-07/ai-risks_d1named-entity-recognition_edpb-spe-programme_en.pdf Isabel Barberá, *AI Possible Risks & 緩和 - Optical Character Recognition*, September 2023, https://www.edpb.europa.eu/system/files/2024-06/ai-risks_d2optical-character-recognition_edpb-spe-programme_en_2.pdf

⁹ISO 31000:2018 に基づく。

¹⁰測定が困難な場合が多い定量的リスクアセスメントとは対照的である。

¹¹イサベル・バルベラ, *AI の潜在的なリスクと緩和 - 固有表現認識*, 2023 年 9 月, https://www.edpb.europa.eu/system/files/2024-07/ai-risks_d1named-entity-recognition_edpb-spe-programme_en.pdf

発生可能性	非常に高い	中	高	非常に高い	非常に高い
	高い	低	高い	非常に高い	非常に高い
	低	低い	中	高い	非常に高い
	ありえない	低い	低い	中	非常に高い
		非常に限定的	限定的	顕著	非常に重大
深刻度					

図 2 : リスクの定性的マトリクス¹²

リスク評価はリスクアセスメントの最終段階であり、リスク分析の結果を組織のリスク規準（リスク許容度やリスク許容範囲など）と比較し、各リスクが許容可能か、あるいは対応が必要かを判断する。この評価の結果は、リスクの深刻度と組織の目標に応じて、リスクを回避、緩和、移転、または受容するかを組織が決定する助けとなる。

リスクアセスメントの次はリスク対応であり、その目的はこれらのリスクを効果的に緩和するための対策を選択し実施することだ。これは反復的なプロセスであり、いくつかの重要なステップを含む。まず、リスク対応策を策定し選択する。次に、特定されたリスクに対処するための計画を策定し実施する。実施後、対策の有効性を評価し、リスクが十分に緩和されたかを判断する。残存リスクが許容可能と判断された場合、追加措置は不要である。しかし、リスクが依然として許容できない場合、さらなる低減のための追加措置が講じられる。

本ガイダンスは、リスクマネジメントプロセスのうち**特定二つの側面**、すなわち**リスク識別**と**リスク対応**に焦点を当てる。**リスク分析**と**リスク評価**の側面は、特定の処理状況に依存しすぎるため、各組織が自らのリスク規準に沿ってアセスメントを行う方が適切である。これは、EUI が使用する予定の AI システムごとに徹底的な分析を行い、リスクの発生可能性と影響を評価するとともに、それらに対処するための緩和措置と残存リスクを決定すべきことを意味する。¹³ この分析の結果、EUI が計画中の AI システムがもたらすリスクを合理的な手段で軽減できないという結論に至り、組織のニーズを満たす別の解決策を見出す必要がある場合もある。その場合、EUI は EUDPR 第 40 条(1)に基づき、事前に欧州データ保護監督官（EDPS）に相談しなければならない。

¹²同上

¹³

3 AI ライフサイクルの理解

3.1 AI システムの定義

本ガイダンスの目的上、AI システムは規則 2024/1689 第 3 条(1) (AI 法) 第 3 条(1)の定義に基づき、「自律性の異なるレベルで動作するよう設計された機械ベースのシステムであり、展開後に適応性を示す可能性があり、明示的または暗黙的な目的のために、受け取った入力から、物理的または仮想的環境に影響を与える予測、コンテンツ、推奨事項、決定などの出力を生成する方法を推論する」と理解される。¹⁴ ただし、AI 法には「AI モデル」の定義は含まれていない。¹⁵ AI システムと AI モデルという用語は、しばしば同義語であるかのように使われるが、実際には異なる。

AI モデルとは、訓練用個人データに潜むパターンをパラメータの集合で表現した数学的モデルである。¹⁶ AI モデルは AI システムの必須要素だが、単独では AI システムを構成しない。機能し、ユーザーや仮想/物理環境と対話するには、常に他のソフトウェアコンポーネントが必要だからだ。実際、AI システムは複数の AI モデルで構成される場合がある。例えば音声翻訳 AI システムは、音声データをテキストに変換する第一モデル、テキストを言語間翻訳する第二モデル、翻訳テキストから音声データを出力する第三モデルで構成され得る。

3.2 AI システムのライフサイクル

リスクは AI システムの開発ライフサイクルの様々な段階で発生しうる。したがって、従来の（非 AI システム向けの）開発ライフサイクルと比較した、AI システム開発ライフサイクルの特性を理解する必要がある。¹⁷ 開発ライフサイクルの各段階では異なるリスクが発生しうる（第 4 節および第 5 節参照）。AI 開発ライフサイクルは通常、図 3 に詳細を示すステップで構成される。¹⁸

- 1 構想/分析：**この初期段階では、AI システムが解決すべき課題を明確に定義し、AI モデルのアーキテクチャを選択する。
- 2 データ取得と準備：**必要なトレーニング用個人データは、AI システムの目的によって異なる。例えば、画像を処理する AI システムには、画像データがトレーニング用個人データとして必要となる。これらの画像は様々なソース（インターネット、私有データベースなど）から取得できる。特定の AI システムに投入するトレーニング用個人データは、使用前にフォーマット化され、適用される品質および法的要件に対してチェックされ、正規化される必要がある。
- 3 開発：**AI システムは、事前に定義された限定的な機能を遂行するようプログラミングおよび訓練される。この段階には、適切なアルゴリズムの選択、準備されたデータを用いた AI システムの訓練、テスト（AI システムが機能しバグがないか確認するため）、および性能向上のためのハイパーパラメータ（学習率など）の調

¹⁴ 欧州議会及び理事会規則 (EU) 2024/1689 (2024 年 6 月 13 日) は、人工知能に関する調和された規則を定め、規則 (EC) No 300/2008、(EU) No 167/2013、(EU) No 168/2013、(EU) 2018/858、(EU) 2018/1139、(EU) 2019/2144、並びに指令 2014/90/EU、(EU) 2016/797、(EU) 2020/1828 を改正する (人工知能法) EC [2024] OJ L2024/1689 <http://data.europa.eu/eli/reg/2024/1689/oj>。

¹⁵ 汎用 AI モデルは「大規模な自己教師あり学習により大量のデータで訓練された AI モデルを含み、上市方法に関わらず広範な異なるタスクを適切に遂行できる顕著な汎用性を示し、様々な下流システムやアプリケーションに統合可能な AI モデル」と定義される (AI 法第 3 条(63))。

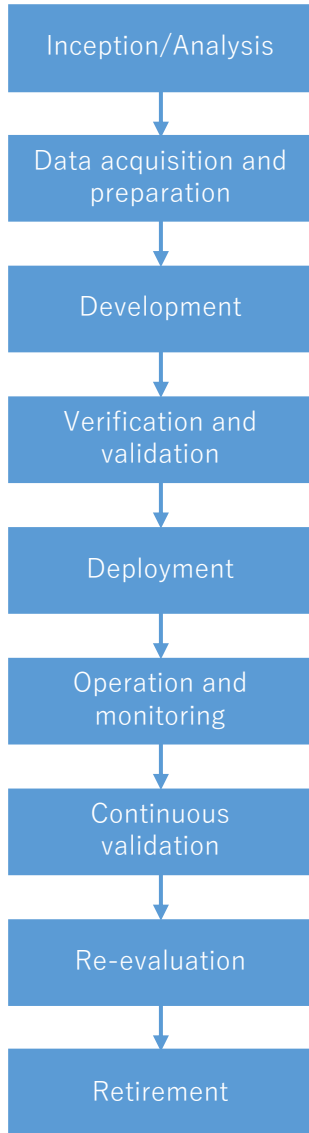
¹⁶ ISO/IEC 22989:2022 は AI モデルを「システム、事業者、現象、プロセス、またはデータの物理的、数学的、あるいはその他の論理的表現」と定義する。

¹⁷ ISO/IEC 15288、ISO/IEC 12207、ISO 24748:2024 を参照のこと。

¹⁸ 詳細は ISO 22989:2022 のプロバイダに記載されている。

整が含まれる。AIシステムの構築には、「ライブラリ」（調達可能）、取得済みの事前学習済みモデル、内部開発を組み合わせることがある。

4 **検証と妥当性確認**：開発段階の後、AIシステムは厳密に検証される（「正しく製品を構築しているか？」）と妥当性確認される（「正しい製品を構築しているか？」）。これにより、構想段階で設定された機能要件と非機能要件を満たしていることが保証される。これには、テストデータセットと検証データセットを用いて、AIシステムの統計的精度、頑健性、汎化能力を確認することが含まれる。この段階でAIモデルに関連する問題が発生した場合、AIシステムを再学習させることで対処する。



5 **展開**：AIシステムは最終環境（エンドユーザーデバイス、サーバー、自動車など）に展開される。

6 **運用と監視**：展開後、AIシステムはユーザーによって運用され、期待通りに動作することを確認するための継続的な監視が必要となる。これには、パフォーマンスの追跡、新たな要件を満たすためのAIシステムの更新、フィードバックに基づく改良が含まれる。

7 **継続的妥当性確認**：AIシステムが継続的学習を利用する場合、¹⁹運用・監視フェーズは継続的妥当性確認フェーズへと拡張される。このフェーズでは、システムが本番環境で稼働している間も継続的にトレーニングが行われる。システムの性能はテストデータを用いて定期的に評価され、正常な動作が確認される。さらに、テストデータは現在の本番データをより正確に反映させるため、定期的に更新が必要となる場合があり、これによりAIシステムの能力をより正確に評価できる。

8 **再評価**：運用・監視フェーズおよび継続的妥当性確認を経て、AIシステムの性能結果に基づき再評価が必要となる場合がある。システムの運用結果は徹底的に分析され、AIシステムに関連する識別されたリスクと比較されるべきである。これにより、識別されたリスクが適切に緩和されたかどうかを検証する。この段階において、これまで識別されていなかったリスクが現れる可能性がある。これらは、セクション2で提示したリスクマネジメントプロセスの次のサイクルで対処する必要がある。

9 **廃止**：AIシステムは、不要になった場合やより高度なソリューションに置き換えられた場合、責任を持って効率的に廃止すべきである。

図3：AI開発ライフサイクル

3.3 AIシステムの調達

さらに多くの場合、AIシステム構築には外部専門家の協力や、機能・データ・セキュリティなどをカバーする商用製品の導入が必要となる。こうしたケースでは、調達段階において既に、組織に望ましくないリスクをもたらすソリューションに予算を実際に投入する前に、リスク評価を行うことが重要である。

こうしたケースでは、複数のフェーズのうちの1つがAI外部プロバイダ（AIシステム全体または一部を提供する主体）に割り当てられる。AIシステムの調達に関わるEUI職員は、これらのAIシステムの展開に関わるEUI

¹⁹再訓練を必要とせず、新たなデータから学習することで、時間の経過とともにその性能を適応・改善する能力。

職員（例：IT エンジニア、IT プロジェクトマネージャー、データ保護責任者またはデータ保護コーディネーター）と連携し、入札の技術的要件策定、適切な製品の選定、AI システムの実装実行を行う必要がある。

調達した AI システムを既存インフラに統合するには、いくつかのアプローチが考えられる。規則 2024/2509²⁰ に基づき、EUI が従うプロセスは以下の通りだ：

1. 公表と透明性（第 163 条）：健全な財務管理、透明性、平等な取扱いの原則を確保する。
2. 入札の募集（第 167 条～第 169 条）：必要な仕様を全て明記した公開入札を開始する。入札仕様書には、計画された AI システムを調達する入札者の能力及び関連する技術的・手続き的な品質保証に関する要件を含める。要求すべき情報の詳細は第 5.3 節に記述されている。
3. 選定及び落札規準（第 170 条）：予め定めた規準（例：価格、品質、持続可能性）に基づき提案を評価する。²¹
4. 履行（第 175 条）：実施状況を監視し、遵守を確保する。

この場合、「実行段階」は第 3.2 節で示した段階の一部と同様の段階で構成される。具体的には：

5.1 検証と妥当性確認

5.2 展開

5.3 運用と監視

5.4 継続的妥当性確認

5.5 再評価

5.6 廃止

AI システムの開発と同様に、調達ライフサイクルの各段階では異なるリスクが発生する可能性がある（第 4 章および第 5 章参照）。各リスクについて、リスクが顕在化する可能性のある段階が示されている（青色のボックス）。該当する場合には、示された各段階に対応する緩和策を策定すべきである。

²⁰欧州議会及び理事会による 2024 年 9 月 23 日付規則（EU、Euratom）2024/2509「連合の一般予算に適用される財務規則に関する規則（改正）」。<https://eur-lex.europa.eu/legalcontent/en/TXT/?uri=CELEX%3A32024R2509>

²¹説明責任の原則に基づき、第 4 条及び第 5 条に列挙された懸念事項に関する確認は、管理者の責任で行うものである。

4 解釈可能性と説明可能性は不可欠な条件

AI システムの調達、開発、展開においては、**解釈可能性と説明可能性**が横断的な課題となる。したがって、これらは EU データ保護規則（EUDPR）に基づく義務を確実に遵守するため、AI システムにおけるパーソナルデータの処理のデータ管理者として機能する EUI にとって前提条件である。ただし、解釈可能性と説明可能性は透明性と混同すべきではない。前者の二つの概念は、管理者が AI システムの機能を理解している程度を指す。後者は、データ管理者がデータ対象者に意味のある情報を提供する義務を指す。解釈可能性と説明可能性は、データ管理者がその情報をプロバイダするための手段ではあるが、それ自体では透明性の基準を満たすには不十分である。本稿では、以下に定義する前者の二つの概念について論じる。

解釈可能性とは、特定の「ブラックボックス」モデルや意思決定が人間に理解される度合いを指す。これは AI モデルがどのように意思決定を行うかを把握する能力に相当する。解釈可能なモデルは透明性を持って動作し、入力と出力の関連性を明らかにする。アルゴリズムが解釈可能であれば、人間はその動作を明確かつ理解しやすく説明できる。このため解釈可能性は、ユーザーが AI モデルを理解し信頼できることを保証する上で極めて重要である。

例えば、線形回帰²²を用いて不動産価格を推定する AI モデルが「価格 = 100,000 + (50 × 床面積) + (10,000 × 部屋数) + (30,000 × 郵便番号スコア)」という形式であれば、計算過程が明確に理解できるため、高い解釈可能性を持つと言える。

AI における**説明可能性とは**、特定のモデル予測や決定に対して明確で一貫性のある説明を提供することに焦点を当てる。これは、エンドユーザーが理解できる形で AI モデルがどのように決定を下すかを明らかにする能力を指す。説明可能なモデルは、その出力に対して明確で直感的な説明を提供し、ユーザーが特定の結果の背景にある理由を理解するのを助ける。本質的に、説明可能性はアルゴリズムが特定の決定に至った理由と、その決定を正当化できる方法を強調する。説明可能性には、モデル自体が本質的に解釈可能でなくとも、特徴量が予測に与える影響を要約または可視化する事後分析技術が含まれる場合がある。

例えば、胸部 X 線画像から肺炎を診断するために用いられる畳み込みニューラルネットワーク（CNN）²³ は、モデル内部の複雑な仕組みゆえに本質的に解釈可能ではない。しかし、LIME（Local Interpretable Model-Agnostic Explanations）²⁴ といった説明可能性ツールを用いて、モデルが判断を下す際に X 線画像のどの部分に注目したかを示すヒートマップを生成できる。この説明により、放射線科医はモデル自体がブラックボックスであっても、その推論プロセスを理解できる。

解釈可能性と説明可能性の違いは、前者が AI モデルの内部構造の理解に関わるのに対し、後者はそれらのモデルが下した決定の説明に焦点を当てる点にある。深層ニューラルネットワークのような複雑な AI モデルは、その複雑な構造や異なる構成要素間の相互作用のため、解釈が困難な場合がある。そのような場合、モデル理解よりも決定の説明を優先する説明可能性の方が現実的である。最後に、解釈可能性は通常 AI 専門家や研究者を対象とするのに対し、説明可能性はエンドユーザーへのモデル決定の伝達に焦点を当てる。したがって説明可能性には、より簡潔で直感的な情報提示が求められる。これは以下の点を保証するために必要である：

- 組織が AI システムが期待通りに動作すると信頼できること；

²²平易な言葉で言えば、線形回帰は入力変数と出力変数の関係を推定するモデルである。

²³データ内のパターンや特徴を検知するためにスライディングフィルターを使用する深層学習モデルの一種だ。

²⁴LIME は入力データを乱し、予測値の変化を観察し、予測値周辺で局所的に解釈可能なモデルを学習することで機能する。これにより、人間が理解しやすいシンプルな AI モデルの構築が可能となる。Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. 「なぜあなたを信頼すべきか？：あらゆる分類器の予測を説明するために」、2016 年 8 月 13 日, <https://doi.org/10.1145/2939672.2939778>

- モデルの誤りやバイアスが容易に識別できること；
- 組織が AI システムの誤用を検知できること；
- 意思決定規準が組織の目標に沿っていること；
- AI システムが監査可能であること。

4.1.1 リスク 1：解釈不能または説明不能な AI システム

4.1.1.1 説明

解釈不能または説明不能な AI システムは重大なリスクをもたらす。これらは「ブラックボックス」として動作し、内部の仕組みや意思決定プロセスが人間ユーザーには不透明なままであるため、特定の出力や決定がどのように、なぜ生成されたのかを理解することが困難になる。

このリスクは、AI システムライフサイクルの以下の段階に適用されることに注意すること：

- 選定（AI システムの調達時）
- 検証と妥当性確認（AI システムの開発と調達の両方）
- 運用と監視（AI システムの開発と調達の両方）
- 継続的妥当性確認（AI システムの開発と調達の両方）
- 再評価（AI システムの開発と調達の両方）

4.1.1.2 可能な対策

このリスクに対する可能な対策は以下の通りである：

1. 文書化：適切な文書はドラフトを作成すべきであり、以下を含む：

- a. 使用した AI アーキテクチャの種類（決定木、ニューラルネットワークなど）とその特性（使用した AI アルゴリズムの種類に関する詳細）、およびこの種のモデルとアルゴリズムの選定理由の説明。
- b. トレーニング用個人データの出所と、当該活動に適している理由の詳細。
- c. AI システムがどのように動作し、データ内で識別可能な異なるグループ間でどの程度の精度を示すかについての情報。
- d. 潜在的なバイアスの説明と、全体的な品質向上およびバイアス発生確率低減のために講じた差異の解消策と対策。
- e. システムの限界の説明。システムが実行可能なこと、不可能なことに対する期待値を明確にする。

この文書は、AI の動作内容とその仕組みを説明する出発点である。管理者はこの文書を読み、AI システムの動作に関する情報を得て、自身のパーソナルデータの処理に関して AI システムの動作が公平かどうかを確認できる。文書はユーザーにとって関連性があり、有用で、理解しやすいものであるべきだ。

2. 説明可能性のための手法（LIME や SHAP（Shapley Additive Explanations）など）の検討 ²⁵

²⁵協調ゲーム理論に基づく手法で、AI モデル内の各特徴量に値を割り当てる。その後、特定のインスタンスに対する予測において、全ての可能な特徴量組み合わせを考慮しながら、各特徴量の寄与度を算出する。この技術は特徴量の重要性を統一的に測定し、AI モデルの決定を説明す

3. 統計分析：²⁶ AI の出力を統計的に分析し、結果の根拠（あるいは根拠の欠如）を説明する。

るのに役立つ。参照：Lundberg, S. M., & Lee, S. I., *A unified approach to interpreting model predictions*. *Advances in neural information processing systems*, 2017 年 11 月 25 日, <https://arxiv.org/pdf/1705.07874>

²⁶ICO, “Task 4: Translate the rationale of your system’s results into useable and easily understandable reasons”, ICO website, 2025 年 8 月 6 日, <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificialintelligence/explaining-decisions-made-with-artificial-intelligence/part-2-explaining-ai-in-practice/task-4-translate/>

5 主要なデータ保護原則に関連するリスク

5.1 公正性の原則

EUDPR は公平性を明示的に定義していないが、これは EU 基本権憲章（憲章）第 8 条(2)に規定されるデータ保護法の一般原則を構成する。²⁷データ管理者による公正な処理の原則遵守義務は、EUDPR 第 4 条(1)(a)および業務上の個人データ処理に関する事案においては第 71 条(1)(a)に規定されている。公正性の原則は、合法性および透明性の原則と本質的に関連しつつも、独立した意味を持ち、他のデータ保護原則への遵守とは無関係に単独で評価される。²⁸

欧州データ保護委員会（EDPB）は、公正性は包括的な原則であり、個人データがデータ対象者に対して不当に不利益をもたらす、違法に差別的である、予期せぬ、または誤解を招くような方法で処理されてはならないことを要求すると明確にしている。²⁹ 処理が公正であるためには、データ対象者が自身から収集された個人データの使用法とその処理の影響について明確に理解している必要がある。公平性は、処理がデータ対象者の合理的な期待を超えないよう、データ管理者側に透明性を求めるものである。

しかし公平性は、透明性の要件を超える義務を課す。公正な処理義務を遵守するためには、処理が関係者の集団的・個人的な利益及び基本的権利に与える影響の評価を行い、個人データが彼らに不当な悪影響を及ぼす可能性のある方法で利用されるべきではない。³⁰ また、データ保護枠組み下での権利と利益の均衡を図るため、データ収集・処理及び権利行使に関する手続き上の保護措置をデータ管理者が実施することも要求される。この意味で、公平性は良き行政の原則とも関連する。同原則は EU 機関に対し、人々の事務を「公平かつ公正に、かつ合理的な期間内に」処理することを求めている（基本権憲章第 41 条）。

このように、公正な処理の原則はデータ保護枠組み全体を支える基盤であり、管理者とデータ対象者の間の権力格差に対処し、そのような格差による悪影響を相殺し、データ対象者の権利が効果的に行使されることを確保しようとするものである。これは、個人データが AI システムで処理される場合に特に重要である。AI システムの機能や影響は、データ管理者自身でさえ把握が困難な場合があるからだ。複雑な AI システムに依存して個人に関する決定を下すことは、EUI がこれらの決定を正当化し、その根拠を示すこともより困難にする。

この文脈において公平性の原則に違反する可能性のある重要なリスクの一つは、バイアスの存在である。欧州データ保護監督官（EDPS）の「生成的人工知能と個人データ保護に関する指針」で既に指摘されているように、「人工知能ソリューションは既存の人間のバイアスを増幅し、新たなバイアスを取り込む可能性がある」。³¹ したが

²⁷基本権憲章第 8 条(2)は「個人データは、特定された目的のために公正に処理され、かつ、関係者の同意または法律で定められたその他の正当な根拠に基づいて処理されなければならない」と規定している。

²⁸欧州データ保護会議、ガイドライン 03/2022「ソーシャルメディアプラットフォームインターフェースにおける欺瞞的な設計パターン：その認識と回避方法」、2023 年 2 月 14 日、第 9 項、https://www.edpb.europa.eu/our-work-tools/ourdocuments/guidelines/guidelines-032022-deceptive-design-patterns-social-media_en。欧州データ保護会議、アイルランド監督当局が提出した *TikTok Technology Limited* に関する紛争に関する拘束力のある決定 2/2023（GDPR 第 65 条）、2023 年 8 月 2 日、パラグラフ 100-107 も参照のこと。https://www.edpb.europa.eu/our-work-tools/ourdocuments/binding-decision-board-art65/binding-decision-22023-dispute-submitted_en

²⁹欧州データ保護会議、設計時及びデフォルト時におけるデータ保護に関するガイドライン、2020 年 10 月 20 日、第 69 項、https://www.edpb.europa.eu/our-work-tools/ourdocuments/guidelines/guidelines-42019-article-25-data-protectiondesign-and_en

³⁰欧州データ保護会議、ガイドライン 2/2019「データ対象者へのオンラインサービス提供における GDPR 第 6 条(1)(b)に基づくパーソナルデータの処理」、バージョン 2.0、2019 年 10 月 8 日、第 12 項、https://www.edpb.europa.eu/our-work-tools/ourdocuments/guidelines/guidelines-22019-processing-personal-data-underarticle-61b_en

³¹欧州データ保護監督官、生成的 AI と EUDPR。生成的 AI システム利用時のデータ保護コンプライアンス確保に向けた初の EDPS 指針（バージョン 2）、2025 年 10 月 28 日、第 13 節（p31）、https://www.edps.europa.eu/system/files/2025-10/25-10_28_revised_genai_orientations_en.pdf。

って、AI システムに依存する EUI は、その訓練に使用されたデータセットに含まれるバイアスを複製するリスクも抱えており、これが差別的な結果を招くことになる。こうしたシステムが個人に影響を与える決定を行うために使用される場合、この問題は特に深刻である。

本ガイドンスにおいて、公平性の原則とは、管理者がこうしたバイアスを識別・測定・緩和することを要求するものと理解される。³² 処理の公平性を確保するため、パーソナルデータの処理を伴う AI システム（特に個人に関する決定の支援・実行に用いられるもの）を調達・開発・展開するデータ管理者は、こうしたバイアスを識別・測定し、あらゆる形態の差別的結果を防止・是正するための技術的・組織的措置を実施すべきである。

AI におけるバイアスの定義は、単一で普遍的に受け入れられているものではない。しかし、一般的に、特定の集団や入力タイプを優遇または差別する、不公平で偏見に満ちた、あるいは体系的に誤った結果を生むことを指す。³³

AI システムにおけるバイアスは、偏見的な視点（例：警察活動において人種や民族集団に不当に焦点を当てる）や不公平な選好（例：高所得郵便番号地域からの融資申請を不釣り合いに承認する）を組み込んだ結果を生み出す可能性がある。

AI におけるバイアスの根本原因としては以下が挙げられる：³⁴

- アルゴリズムバイアス：AI システム自体の設計が偏った結果を生むことがある。特定の AI モデルやアルゴリズムを採用する決定、および AI システム開発に特定の情報を組み込むことが、不公平な結果につながる可能性がある。
- トレーニング用個人データ：AI システムはトレーニング用個人データから学習する。効果的に AI システムを訓練するためには、通常、大量のトレーニング用個人データが必要となる。³⁵ トレーニング用個人データにバイアスが存在する場合、AI システムはそのバイアスを知覚し、同様にバイアスのかかった結果を生成する。例えば、男性が歴史的に特定の職種を占めてきた場合、歴史的データで訓練された AI システムは、そのような歴史的バイアス保持し、男性がそれらの職種に最も適した候補者であると学習する可能性がある。特定の人口統計に属する個人の顔で訓練された顔認識システムも同様に、訓練用個人データで過小評価または未評価の個人の顔に直面した場合、高い統計的精度を達成するのに苦勞する。
- その他人間によるバイアス：開発者や AI システムの訓練・運用責任者は、意識的・無意識的なバイアス AI システムの設計や実装に持ち込む可能性がある。例えば、訓練プロセスの一部で人間が結果をレビューする場合、個人が明示的・無意識的なバイアスに基づいて結果を拒否または承認する選択をするかもしれない。

5.1.1 リスク 1：訓練用個人データの品質不足によるバイアス

5.1.1.1 説明

³²AI 法は、バイアスや代表者データセットに関する同様の問題に対処するため、第 10 条に具体的な要件を定めている。

³³IBM, 「AI バイアスとは何か?」, IBM ウェブサイト, 2025 年 8 月 6 日, <https://www.ibm.com/think/topics/ai-bias>

³⁴バイアスの他の発生源も存在する。例えば以下を参照のこと。Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). 機械学習におけるバイアスと公平性に関する調査。ACM コンピューティングサーベイ (CSUR)、2022 年、<https://arxiv.org/abs/1908.09635>

³⁵例えば、画像分類には数百万枚の画像が必要となる場合がある。大規模言語モデルは通常、トークンと呼ばれる数十億から数兆のテキストで訓練される。

AI システムは「ゴミを入れればゴミが出る」という原理で動作するため、訓練段階では質の高いデータが必要だ。³⁶ 不正確または不完全な訓練用個人データセットは、AI システムからの誤った出力につながる。例えば、誤ったラベル付け情報を含むデータセットで画像認識プログラムを訓練すると、プログラムはそれらの誤りを再現し、誤ったラベルを提供する結果となる。³⁷

このリスクは、AI システムのライフサイクルにおける以下の段階に適用される：

- データ取得と準備（AI システム開発）
- 検証と妥当性確認（AI システムの開発と調達の両方）
- 再評価（AI システムの開発と調達の両方）

5.1.1.2 可能な対策

データ品質リスクに対処するには、データ取得・準備段階で用いられる個人データが、AI システムが使用される対象集団を多様かつ代表者（すなわち、偏りのない）に反映していることを確保することが不可欠である。これには、幅広い情報源からのデータ収集と、過小評価されているグループの包含に向けた取り組みが必要だ。個人データセットの定期的な監査と更新は、その関連性と包括性を維持するのに役立つ。

1. 訓練用個人データセットの品質保証方針を定義する。その内容は以下の通りである：
 - a. 収集するデータの種類とデータ取得方法を記述する。
 - b. トレーニング用の個人データの準備過程で実施する手順を説明する（クリーニング、³⁸ ラベリング、³⁹ 正規化とスケーリング、⁴⁰ 分割⁴¹）。
 - c. 品質規準と測定方法の定義を示す。
 - d. 品質閾値を定義する。
2. ポリシーに従い、データセットをサンプリングし、合意された品質閾値に対して測定・評価する、トレーニング用個人データセットの評価手順を定義し実施する。
3. AI システムのトレーニング個人データについて、データ品質を確認するための定期的な監査を実施する。
4. 統計的手法を用いて外れ値を検知する。外れ値は検証が必要であり、有効な値（訓練用個人データに残すべき）か、誤った値（削除すべき）かを判断する。例えば、生年月日を扱う訓練用個人データにおいて、100 歳を超える年齢を示す日付は精査すべきである。これらの誤りは識別後、必要に応じて以下の方法で修正できる：手動介入、統計的手法（他の値の平均値や中央値を計算して推定値を算出する手法など）、回帰分析（線形回帰、多項式回帰など）、あるいは K 近傍法のような高度な統計手法（類似した訓練用個人データを用いて異常値を補正する手法）を用いる。必要と判断された場合は訓練用個人データから削除することも可能である。

³⁶この原則は、あらゆるシステムにおいて、出力の質は入力質によって決まるという考え方である。

³⁷意味のあるタグや注釈を、画像・テキスト・音声などの生データに付与するプロセス。これにより正しい出力やカテゴリーを示し、教師あり機械学習モデルを訓練して正確な予測を行う。

³⁸データセット内の誤り、重複、無関係な情報（欠損値、外れ値、矛盾など）を除去または修正する。

³⁹AI が正しい関連性を学習するよう、ラベルの正確性を検証する。

⁴⁰特徴量を正規化してデータセットを標準化し、均一性を確保するとともに、特定の属性がモデルに過度に影響するのを防ぐ。

⁴¹データセットを訓練用、妥当性確認用、テスト用に分割し、モデルが見たことのないデータにも適切に一般化でき、過学習しないようにする。

5. トレーニング用個人データは検証可能であり、情報の妥当性確認およびバイアスのリスク最小化できる。これには、信頼できる情報源からの個人データの取得、利用可能な場合における信頼できる類似情報源との照合、専門知識を持つ担当者による人的レビューの実施、ファジーマッチング技術（個人データの近似度を確認する手法）の活用、統計的手法によるデータのクラスタ検知（潜在的なバイアスの識別）、および個人データの出所記録、正確性⁴² 妥当性確認・トレーサビリティの正当化が含まれる。⁴³
6. 標準化と一貫性を確認し、全てのトレーニング個人データ項目が同一の形式で表現されていることを保証する（例：生年月日は全て DD/MM/YYYY 形式を使用）。

5.1.2 リスク 2：トレーニング個人データにおけるバイアス

5.1.2.1 説明

機械学習モデルは、正確かつ完全なトレーニング用個人データで訓練されていても、出力にバイアスが生じる可能性がある。トレーニング用個人データに関連するバイアスの主な原因は以下の通りである：⁴⁴

- データ収集時のサンプリング誤差は母集団バイアスをもたらす。これは、訓練用個人データがより広範な母集団を代表していない場合に発生する。例えば、主に都市部の病院データを用いて開発された医療 AI システムは、患者層や健康状態が異なる地方では十分な性能を発揮しない可能性がある。多様性や代表者に欠けるデータでは、AI システムが異なる母集団や環境で予測を一般化できないことを意味する。
- 歴史的バイアスとは、社会に存在する既存の偏見や社会技術的問題を指し、バイアス防止技術を用いてもデータに浸透する可能性がある。例えば、歴史的に女性最高経営責任者（CEO）の数は少なく、現在も同様である。CEO の画像を検索する AI は偏り、主に男性 CEO の画像を表示する可能性がある。

このリスクは、AI システムライフサイクルの以下の段階に適用される：

- データ取得と準備（AI システム開発）
- 検証と妥当性確認（AI システムの開発と調達の両方）
- 再評価（AI システムの開発と調達の両方）

5.1.2.2 可能な対策

組織がトレーニング用個人データにアクセスできる場合、トレーニング用個人データ内のバイアスを検知・修正する技術を活用できる。これらの手法には、異なるグループの表現バランスを調整する統計的補正や、検知されたバイアスの影響を緩和するアルゴリズム的介入が含まれる。

1. AI システムが信頼性が高く偏りのない出力を生成するには、代表者のトレーニング個人データが不可欠である。トレーニング個人データが実世界のデータと大きく異なる場合、システムは誤った推論を導く可能性が高い。これには以下が含まれる：

⁴²AI システムの出力結果が、期待される結果または真の結果と一致していること。

⁴³AI システムが与えられた文脈において適切であり、意図した通りに機能しているか。

⁴⁴Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A, A survey on バイアス and fairness in machine learning. ACM computing surveys (CSUR), 2022, <https://arxiv.org/abs/1908.09635> バイアスの種類については、Mikołajczyk-Bareła, A., & Grochowski, A survey on bias in machine learning research, 2023, <https://arxiv.org/abs/2308.11254> に記述されている。

- a. 分布の一致性：訓練データと対象入力データの両方における主要特徴量の統計的分布が類似しているかを検証する。ヒストグラム、要約統計量、クラスタリングなどのツールで差異や欠落を識別できる。
 - b. 多様性と網羅性：入力データの範囲と多様性が、訓練用個人データによって適切に表現されているかを分析する。代表者である訓練用個人データは、実世界の運用データに存在する全ての意味のあるシナリオ、クラス、バリエーションを網羅すべきである。
 - c. 妥当性確認と交差検証：想定される運用データを模倣した妥当性確認データセットを使用する。交差検証や訓練/妥当性確認フェーズのローテーションを適用することで、モデル性能が見たことのないデータに対しても一貫性と頑健性を保つかどうかを測定できる。
 - d. 専門家によるレビューとシナリオ検証：専門知識を持つ担当者が主要変数を定義し、訓練個人データに関連する運用上の側面を全て包含しているかを検証すべきだ。クラス不均衡やサンプリングバイアスの有無を確認することも重要である。
2. バイアスフリーな特徴量：⁴⁵ バイアスを導入しにくい特徴量を選択する。人種、性別、社会経済的地位などの機微な属性を直接コード化する特徴量は、正当な理由と慎重な取り扱いがない限り避ける。機微な属性の代理変数として機能する特徴量（例：世帯収入の代理変数としての郵便番号）を検知するよう努める。
 3. 特徴量エンジニアリング：⁴⁶ AI モデルに組み込む特徴量の選択は、AI モデルの挙動に重大な影響を与える。特定のバイアスに基づいて特徴量が選ばれると、AI モデルの予測結果にもそのバイアスが見られる。このプロセスでは、使用する特徴量が関連性があり、意図せずバイアスをもたらさないよう慎重な検討が必要だ。さらに、特徴量をバイアスを軽減する形で変換することも可能である。例えば、特徴量の重み付けや再スケーリングを行うことで、単一の特徴量が AI モデルの結果に過度に影響するのを防げる。⁴⁷
 4. バイアス監査：⁴⁸ AI システムのトレーニング用個人データを定期的に監査し、バイアスがないか確認する。
 5. データに対するバイアス緩和技術：再重み付けなどのデータに対するバイアス緩和技術は、AI システムが使用するデータ内の識別されたバイアスを軽減できる。⁴⁹

5.1.3 リスク 3：トレーニング個人データへの過学習

5.1.3.1 説明

AI システムにおいて過学習とは、訓練用個人データの詳細やノイズを学習しすぎて、訓練データそのものを再現する傾向が生じ、新規の未見データに対する AI システムの性能に悪影響を及ぼす現象を指す。言い換えれば、

⁴⁵特徴量とは、データセット内のデータポイントの属性である（例：年齢、郵便番号、身長）。

⁴⁶特徴量エンジニアリングとは、機械学習モデルの性能と解釈性を向上させるため、生データから関連する変数を選択・変換・生成するプロセスである。

⁴⁷再スケーリング：データ値の範囲や分布を再調整するプロセス。多くの場合、機械学習モデルにデータを投入する前に行われる。

再重み付け：AI モデル、特に機械学習やニューラルネットワークにおいて、様々な入力、特徴量、または接続に対して数値的価値、すなわち「重み」を再割り当てする行為である。

⁴⁸IBM、「AI Fairness 360 の紹介」、IBM リサーチウェブサイト、2025 年 8 月 6 日、<https://research.ibm.com/blog/ai-fairness-360Aequitas>、「機械学習のためのバイアスと公平性監査ツールキット」、Aequitas ウェブサイト、2025 年 8 月 6 日。
<https://dssg.github.io/aequitas/>

⁴⁹過小評価されているグループからのサンプルに異なる重みを割り当て、モデルへの公平な影響を確保する。詳細はエマヌエル・クラサナキス、エレフテリオス・スピロミトリス＝シウフィス、シメオン・パバドプロス、ヤニス・コンパティアリス著『公平性を考慮した分類におけるバイアス緩和のための適応型センシティブ再重み付け』を参照。2018 年世界ワイドウェブ会議（WWW '18）論文集収録。国際ワールドワイドウェブ会議運営委員会、ジュネーブ州、スイス、2018 年 4 月 23 日、<https://doi.org/10.1145/3178876.3186133>

モデルが訓練用個人データの特定の詳細やノイズを徹底的に学習し、事実上個人データを暗記してしまう場合に過学習が発生する。これは、AI システムが過度に複雑化し、訓練用個人データセット固有のパターンを捉えるが他のデータにはうまく一般化できない場合、あるいは訓練用個人データセットが適切な最小サイズを満たしていない場合に生じる。

このリスクは、AI システムのライフサイクルにおける以下の段階に適用される：

- データ収集と前処理（AI システム開発）
- 検証と妥当性確認（AI システムの開発と調達の両方）
- 運用と監視（AI システムの開発と調達の両方）
- 継続的妥当性確認（AI システムの開発と調達の両方）
- 再評価（AI システムの開発と調達の両方）

5.1.3.2 可能な対策

このリスクに対する可能な対策は以下の通りである：⁵⁰

1. 早期停止：早期停止とは、妥当性確認データセットにおける AI モデルの性能が低下し始めた時点で、トレーニングプロセスを直ちに停止させる手法である。これは、トレーニング用個人データへの過学習の可能性を示している。
2. 簡素化：AI モデルの簡素化も過学習の緩和という実用的な手法である。これは、より少ない関連性の高い特徴量を選択したり、重要度の低いパラメータやニューロンを除去して AI モデルを剪定したりすることを含む。⁵¹ AI モデルの複雑性を低減することで、訓練用個人データ内のノイズへの過学習の可能性が低下し、一般化能力が向上する。
3. 正則化手法：正則化手法とは、機械学習においてモデルの複雑さにペナルティを加えることで過学習を防ぎ、新規データへの汎化性能を確保する手法である。最も一般的な手法は L1 正則化と L2 正則化だ。L1 正則化（Lasso）はモデルの重みの絶対値に基づくペナルティを加え、一部の重みをゼロに近づけることで疎さを促進する。これは事実上の特徴量選択となる。L2 正則化（リッジ）は重みの二乗に基づくペナルティを加え、重みをゼロ方向へ収縮させるが完全に除去せず、滑らかさを促進し特徴を捨てずに過学習を防ぐ。エラスティックネットは L1 と L2 正則化を組み合わせ、特徴選択と重みの安定性のバランスを取る。これは特に相関の高い特徴を扱う場合に有用である。
4. ドロップアウト：ドロップアウトはニューラルネットワークで一般的に用いられる別の正則化手法である。訓練中にランダムに選択されたニューロンを無視する。これにより AI モデルが特定のニューロンに過度に依存するのを防ぎ、ネットワークがより頑健な特徴を学習するよう促す。

5.1.4 リスク 4：アルゴリズムバイアス

5.1.4.1 説明

⁵⁰Ying, Xue. (2019), 過学習とその解決策の概観. *Journal of Physics: Conference Series*, 2019, <https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022/pdf>

⁵¹データが持つ個々の測定可能な特性や特徴であり、モデルが予測や分類を行う際に利用されるものである。Domino.ai, 「機械学習とデータサイエンスにおける特徴量とは何か?」, Domino.ai ウェブサイト, 2025 年 8 月 6 日, <https://domino.ai/data-science-dictionary/feature>

アルゴリズムバイアスとは、入力データや訓練用個人データとは独立して、AI システム自体の設計から生じる偏りを指す。⁵²

アルゴリズムの設計方法もバイアスをもたらす可能性がある。例えば、アルゴリズムの性能を最適化するために使用する数学機能の選択、過学習を防ぐために用いる手法、統計モデルをデータセット全体に適用するか特定のサブグループに適用するかという決定など、いずれもバイアスをもたらす要因となり得る。米国司法制度で再犯率予測に用いられた COMPAS モデル⁽⁵³⁾ は、アフリカ系アメリカ人被告に対するバイアスを有していると判明した。原因の一つは、モデルが特定の特徴量と予測値の間に誤った線形関係を仮定していたことである⁽⁵⁴⁾。さらに、バイアスが生じやすい統計手法の使用もアルゴリズムの結果に影響を与える。こうした設計上の選択はアルゴリズムの判断に影響し、特定の集団を不当に有利または不利にするバイアスの生じた結果を招く可能性がある。

このリスクは、AI システムライフサイクルの以下の段階に適用される：

- 構想／分析（AI システム開発時）
- 検証と妥当性確認（AI システムの開発と調達の両方）
- 継続的妥当性確認（AI システムの開発と調達の両方）
- 再評価（AI システムの開発と調達の両方）

5.1.4.2 可能な対策

公平性を考慮したアルゴリズム、バランスの取れた目的関数、慎重な特徴量エンジニアリング、定期的な監査、透明性、公平性指標⁵⁵、包括的な開発手法を取り入れることが重要だ。これにより、多様な集団において公平に機能する AI システムを構築できる。⁵⁶

1. 公平性を考慮したアルゴリズム：⁵⁷ 公平性の制約を考慮して設計されたアルゴリズムを選択する。例えば Two Naive Bayes⁽⁵⁸⁾ のようなアルゴリズムは、公平性問題に対処するために特別に開発されており、構想・分析段階でバイアスの緩和に役立つ。
2. 目的関数の選択：⁵⁹ アルゴリズムは通常、精度最大化や誤差最小化といった目的関数で定義された特定目標達成のために最適化される。しかし、これらの関数が異なるグループ間の公平性を考慮しない場合、バ

⁵²Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). 機械学習におけるバイアスと公平性に関する調査。ACM コンピューティングサーベイ (CSUR)、2022 年 1 月 25 日、<https://arxiv.org/pdf/1908.09635>

⁵³代替制裁のための矯正犯罪者管理プロファイリング (COMPAS)

⁵⁴シンシア・ルーディン、キャロライン・ワン、ポー・コーカー、「再犯予測における秘密主義と不公平の時代」、2020 年 5 月 31 日、<https://hdr.mitpress.mit.edu/pub/7z10o269/release/7>

⁵⁵メトリクスとは、様々なタスクにおける AI システムの性能と有効性を評価するために用いられる定量的指標である。分類、回帰、クラスタリングなど、異なる AI モデルには異なるメトリクスが必要となる。多くの場合、単一のメトリクスでは性能の全体像を捉えきれないため、AI システムの性能の異なる側面を包括的に評価するために複数のメトリクスを算出することが推奨される。

⁵⁶多様性と包摂性を備えたチームの構築。詳細は Moeed Yusuf 博士『アルゴリズム的正義：コードに潜むバイアス、社会に広がるバイアス』（2024 年）を参照のこと。<https://journalpsa.com/index.php/JPSA/article/view/14/16>、Hernández, E. G, 「組織における人工知能の倫理的かつ包括的な導入に向けて：多次元枠組み」、2024 年 5 月 2 日、<https://arxiv.org/abs/2405.01697>

⁵⁷Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D, 機械学習における公平性を高める介入の比較研究。公平性、説明責任、透明性に関する会議の議事録 (329-338 ページ)、2019 年 1 月、<https://arxiv.org/pdf/1802.04422>

⁵⁸Toon Calders および Sicco Verwer, 差別のない分類のための 3 つのナイーブベイズアプローチ。データマイニングジャーナル、ECML/PKDD からの厳選論文を掲載した特別号、2010 年、<https://www.cs.ru.nl/~sicco/papers/dmkd10.pdf>

⁵⁹予測結果と実際の結果の差を測定するために用いられる数学的定式化であり、モデルの最適化を導くものである

バイアスを導入する可能性がある。例えば、全体的な精度のみを最適化したアルゴリズムは、多数派グループと少数派グループ間の性能格差を無視し、偏った結果を招く恐れがある。

3. バイアス監査：⁶⁰ AIシステムに対して定期的な監査を実施し、バイアスを検証する。これには、異なる人口統計学的グループにおけるアルゴリズムの性能評価、および不均衡の識別とアセスメントが含まれる。
4. 多様なデータでのテスト：アルゴリズムを、遭遇する現実世界のシナリオの多様性を反映した多様なデータセットでテストする。これにより、AIモデルが異なる集団間で公平に動作することを保証するのに役立つ。
5. AIモデルの解釈可能性：⁶¹ 複雑なAIモデルを理解しやすくする解釈可能なAIモデルや技術を用いる。これによりAIモデル内のバイアス要因を識別し対処できる。
6. 問題が機械学習や深層学習以外のアルゴリズムを効果的・効率的に使用することで解決可能か、あるいはニューロシンボリックAIを含む他の手法と統合することで解決可能かを検証する。⁶²

5.1.5 リスク5：解釈バイアス

5.1.5.1 説明

解釈バイアスとは、分析者が事前概念や不完全な理解に影響され、トレーニング用個人データやAIモデルの出力から誤った結論や偏った結論を導く現象である。さらに、性能指標の選択的解釈は公平性に関する問題を覆い隠し、バイアスのかかったAIシステムの展開につながる可能性がある。これはモデルの修正、微調整、再トレーニング時に影響を及ぼす欠陥のある判断を招く恐れがある。

例えば、医療組織がAIを活用した診断ツールを開発し、患者の病歴、症状、検査結果に基づいて特定の疾患の可能性を予測するとしよう。このツールは0から1の確率スコアを出力し、1は疾患の高い可能性を示す。解釈バイアスとは、このツールを使用する医療プロバイダが、出力結果を確率スコアではなく確定診断と誤って解釈する場合を指す。例えばスコア0.8を「患者は確実に疾患を有している」、スコア0.2を「患者は疾患を有していない」と想定する可能性がある。

このリスクはAIシステムライフサイクルの以下の段階に適用される：

- 検証と妥当性確認（AIシステムの開発と調達の両方）
- 運用と監視（AIシステムの開発と調達の両方）
- 再評価（AIシステムの開発と調達の両方）

5.1.5.2 可能な対策

このリスクに対する可能な対策は以下の通りである：

1. 多様なチームの参画：データサイエンティスト、ドメインエキスパート、ステークホルダーからなる多様なチームを参画させることで、個人データのトレーニングやAIモデルの出力結果に対して複数の視点を提供できる。

⁶⁰IBM, 「AI Fairness 360 の紹介」, IBM 研究ウェブサイト, 2025 年 8 月 6 日, <https://research.ibm.com/blog/ai-fairness-360Aequitas>, 「機械学習のためのバイアスと公平性監査ツールキット」, Aequitas ウェブサイト, 2025 年 8 月 6 日.<https://dssg.github.io/aequitas/>

⁶¹Carvalho DV, Pereira EM, Cardoso JS., *Machine Learning Interpretability: A Survey on Methods and Metrics*, 2019 年 7 月 26 日, <https://doi.org/10.3390/electronics8080832>

⁶²Wan, Z., Liu, C. K., Yang, H., Li, C., You, H., Fu, Y., ... & Raychowdhury, A. 認知 AI システムに向けて：ニューロシンボリック AI に関する調査と展望, 2024 年 1 月 2 日, <https://arxiv.org/abs/2401.01040>

2. 明確な文書化とコミュニケーション：データソース、特徴量選択、前処理手順、モデリング上の決定事項について、明確かつ包括的な文書を維持することで、分析プロセスの透明性を確保できる。
3. AI モデルの説明可能性技術：SHAP、LIME、特徴量重要度分析などの AI モデルの説明可能性技術を組み込むことで、AI モデルが意思決定を行う仕組みに関する洞察が得られる。⁶³
4. トレーニングと意識向上：AI 開発プロセスに関わるチームメンバーに対し、バイアス、公平性、解釈可能性に関するトレーニングと意識向上を提供することで、解釈可能性バイアスを識別し対処するチームの能力を向上させられる。
5. バイアス監査：⁶⁴ アナリストによる出力解釈の定期的な監査を実施し、バイアスを確認する。

5.2 正確性の原則

5.2.1 EUDPR における正確性の法的意味

EUDPR 第 4 条(1)(d)項によれば、個人データは正確でなければならず、必要に応じて最新の状態に保たれる必要がある。処理目的に照らして不正確な個人データについては、遅滞なく消去または修正されるよう、あらゆる合理的な措置を講じなければならない。EUDPR における正確性とは、データ管理者が個人データ自体が事実関係に関して誤りや誤解を招くものでないことを保証することを要求する。

5.2.2 AI 開発における正確性の統計的意味

データ保護原則における正確性の意味とは異なり、AI の文脈では正確性とは、AI システムが正しい答えを推測した頻度を予測総数で割った性能指標である。

AI における正確性は、入力された個人データの正確性や予測された個人データそのものの正確性を指すのではなく、AI システムの性能を指すものである。

本ガイダンスでは、今後「正確性」という用語は対応するデータ保護原則を指し、「統計的正確性」は AI システムの正確性を指すものとする。

5.2.3 リスク 1：不正確な個人データの出力

5.2.3.1 説明

統計的正確性を評価しないことは、不正確な個人データを生成する AI モデルの展開につながり、データ正確性の原則に対するコンプライアンスリスクを生じさせる可能性がある。AI システムの出力が十分に妥当性確認されない場合、エラーが検出されないままになる。AI モデルが多様であるため、AI モデルの統計的正確性を評価するために使用できる指標も多岐にわたる。例は附属書 1 に示す。

AI モデルは、訓練用個人データや入力データに存在しなかった誤った情報や無意味な情報（個人データを含む）を生成することがある。これは大規模言語モデル（LLM）のようなモデルで発生し、事実を「創作」したり、確信を持って誤った回答を提供したりする可能性がある。こうした「幻覚」は、AI モデルの確率的性質に起因す

⁶³EDPS, 説明可能な人工知能に関する技術情報 #2/2023, 2023 年 11 月 16 日, https://www.edps.europa.eu/dataprotection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en

⁶⁴IBM, 「AI Fairness 360 の紹介」、IBM 研究ウェブサイト、2025 年 8 月 6 日、<https://research.ibm.com/blog/ai-fairness-360Aequitas>、「機械学習のためのバイアスと公平性監査ツールキット」、Aequitas ウェブサイト、2025 年 8 月 6 日、<https://dssg.github.io/aequitas/>

る。AI モデルは、事前に検知・妥当性確認された決定論的ルールに基づく計算を行うのではなく、最も可能性の高い出力を予測しようとするからだ。

さらに、AI システムの統計的精度は、訓練に使用されたデータセットの品質に大きく依存する。⁶⁵ 訓練用個人データが不正確、不完全、またはバイアスがある場合、AI システムは信頼性の低い、あるいは欠陥のある結果を生成する可能性がある。機械学習アルゴリズムは、訓練用データからパターン、行動、関連性を学習するため、このデータに含まれる誤りや虚偽の表現は、AI システムの予測において永続化される可能性がある。なお、良質なデータセットで訓練された AI システムであっても幻覚を起こす可能性があることに留意すべきである。

このリスクは、AI システムライフサイクルの以下の段階に適用される：

- データ取得と準備（AI システム開発）
- 検証と妥当性確認（AI システムの開発と調達の両方）
- 運用と監視（AI システムの開発と調達の両方）
- 再評価（AI システムの開発と調達の両方）

5.2.3.2 可能な対策

このリスクに対する対策としては以下が考えられる：

1. 高品質な個人データによる訓練：正確で信頼性の高い AI モデルを開発するには、高品質な個人データによる訓練が基本となる。AI システムは訓練データから学習するため、データが適切に準備されクリーンであることを保証すれば、モデルの統計的精度を大幅に向上させられる。
2. エッジケース：AI システムは、異常値（アウトライヤー）や敵対的サンプルを用いて検証・妥当性確認を行うべきである。これにより、異常な状況や困難な条件下でのレジリエンスと信頼性を評価できる。⁶⁶
3. 多様で代表者データ：多様なソースからデータを収集し、AI システムが本番環境で遭遇する可能性のある全てのシナリオを代表していることを保証することが不可欠だ。例えば、空港での顔認識用 AI システムを開発する場合、AI システムは代表的な画像（照明条件、表情）で訓練されるべきであり、高解像度で照明の良い正面画像のみに依存すべきではない。
4. バランスの取れたデータセット：分類問題において、各カテゴリーやクラスが均等に表されるようにする。例えば、医療診断モデルでは、特定の結果にバイアスがかからないよう、陽性例と陰性例の双方を十分な数用意する必要がある。
5. ハイパーパラメータ最適化（HPO）：⁶⁷ HPO とは、未見のデータに対するモデルの性能を向上させる最適なハイパーパラメータの組み合わせを見つけることである。ハイパーパラメータとは、機械学習モデルにおける設定値であり、学習前に設定され、モデルの複雑さ、学習率、正則化（大きな重みやパラメータを制限する）といった学習プロセスの側面を制御するが、データから学習されるものではない。

⁶⁵Zhou, Y., Tu, F., Sha, K., Ding, J., & Chen, H., 機械学習のためのデータ品質の次元とツールに関する調査, 2024 年 6 月 28 日, <https://arxiv.org/abs/2406.19614v1>

Budach, Lukas & Feuerpeil, Moritz & Ihde, Nina & Nathansen, Andrea & Noack, Nele & Patzloff, Hendrik & Harmouch, Hazar & Naumann, Felix, データ品質が機械学習モデルの性能に与える影響, 2022 年 7 月, https://www.researchgate.net/publication/362386427_The_Effects_of_Data_Quality_on_ML-Model_Performance

⁶⁶エッジケースとは、極端な（最大または最小の）動作パラメータでのみ発生する問題や状況を指す。

⁶⁷Morales-Hernández, A., Van Nieuwenhuysse, I. & Rojas Gonzalez, S. 機械学習のための多目的ハイパーパラメータ最適化アルゴリズムに関する調査, 2022 年 12 月 24 日, <https://doi.org/10.1007/s10462-022-10359-2>

6. 人間の監視（人間と AI の協働（HAIC）および人間ループ内（HITL））：⁶⁸ AI の意思決定プロセスに人間のレビューを組み込むことで、モデルの予測が二重に確認され、誤りの可能性が低減される。AI システムに対する人間のレビューは、状況、AI アプリケーションの複雑さ、その決定に伴うリスクのレベルに応じて様々な形態をとる。⁶⁹
7. 問題が機械学習や深層学習以外のアルゴリズムを効果的・効率的に使用することで解決可能か、あるいはニューロシンボリック AI を含む他の手法と統合することで解決可能かを検証する。⁷⁰

5.2.4 具体例：データドリフトと入力個人データの品質劣化による不正確な出力

5.2.4.1 説明

データドリフトとは、入力データの統計的特性が時間とともに変化することを指す。⁷¹ これは、ユーザー行動の変化や AI システム運用環境の変化など、様々な要因によって発生する可能性がある。データドリフトは、AI モデルが不正確な予測や判断を行う原因となる。入力データの品質は、ノイズの増加、欠損値、不正確さなどの問題によって劣化することがある。例えば、経済が安定した状況下で申請された融資データを用いて訓練された信用スコアリング AI モデルが、インフレ率や失業率に大幅な変化が生じた経済危機の文脈で使用される場合、これはデータドリフトの代表者である。

このリスクは、AI システムライフサイクルの以下の段階に適用される：

- 運用と監視（AI システムの開発と調達の両方）
- 継続的妥当性確認（AI システムの開発と調達の両方）
- 再評価（AI システムの開発と調達の両方）

5.2.4.2 可能な対策

このリスクに対する可能な対策は以下の通りである：

1. データドリフト検知手法：⁷² 時間の経過に伴うデータ分布の変化を監視するデータドリフト検知手法を導入する。

⁶⁸Fragiadakis, G., Diou, C., Kousiouris, G., & Nikolaidou, M., 「人間と AI の協働評価：レビューと方法論的枠組み」, 2025 年 3 月 7 日, <https://arxiv.org/abs/2407.19098>

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, Liang He, 機械学習におけるヒューマン・イン・ザ・ループの調査, *Future Generation Computer Systems*, 2022 年 10 月, <https://doi.org/10.1016/j.future.2022.05.014>

⁶⁹展開前レビューは、AI システムの開発段階で妥当性確認を行うことを含む。具体的には、トレーニング用個人データの品質妥当性確認、パイアスのテスト、法的枠組みへの準拠確認などだ。ヒューマン・イン・ザ・ループ（HITL）監視は、AI の運用段階における人間の介入を組み込み、人間がリアルタイムで意思決定を監視、修正、承認することを可能にする。意思決定後レビューは、実行後の AI の意思決定を評価し、誤りや改善点を識別することに焦点を当てる。

⁷⁰EDPS, *Techsonar 2025*, 2024 年 11 月 15 日, https://www.edps.europa.eu/data-protection/ourwork/publications/reports/2024-11-15-techsonar-report-2025_en

Wan, Z., Liu, C. K., Yang, H., Li, C., You, H., Fu, Y., ... & Raychowdhury, A, 認知 AI システムに向けて：ニューロシンボリック AI に関する調査と展望, 2024 年 1 月 2 日, <https://arxiv.org/abs/2401.01040>

⁷¹GeeksforGeeks. 機械学習におけるデータドリフト, 2025 年 7 月 23 日, <https://www.geeksforgeeks.org/machine-learning/datadrift-in-machine-learning/>

⁷²Gemaque RN, Costa AFJ, Giusti R, dos Santos EM. 教師なしドリフト検知手法の概観, 2020 年 7 月 21 日, <https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1381>

Andrés L. Suárez-Cetrulo, David Quintana, Alejandro Cervantes, 反復的な概念ドリフトデータストリームに対する機械学習の調査, *Expert Systems with Applications*, 2023 年 3 月 1 日, <https://doi.org/10.1016/j.eswa.2022.118934>

2. データ品質監視：入力データの完全性、統計的正確性、一貫性などの指標を追跡するデータ品質監視システムを導入する。妥当性確認に使用されるデータが高品質を維持していることを確認するため、これらの指標を定期的に見直す。
3. 定期的なモデル再学習：機械学習モデルを定期的に更新し、データパターンの変化に対応させるプロセスである。再学習は固定スケジュール（例：週次・月次）で実施するか、性能低下やドリフト検知をトリガーとして実行する。通常、新規データの収集、前処理、モデル更新、性能向上の妥当性確認という手順を踏む。
4. ユーザーフィードバック：ユーザーが推論における問題や異常を報告できるフィードバックループを構築する。このフィードバックを活用して潜在的なデータ品質問題やドリフトを識別し、AI モデルに必要な調整を行う。

5.2.5 リスク 2: AI システムプロバイダからの不明確な情報

5.2.5.1 説明

事前学習済み AI システムを調達する際のデータ保護リスクを効果的に管理するには、組織は AI プロバイダが採用する開発プロセスの理解に注力すべきだ。

これには、AI システム全体に関する質問を行うとともに、構想/分析、データ取得・準備、開発・検証、妥当性確認の各段階（3.2 節参照）に関連するリスクが適切に管理されたかについて詳細を掘り下げ、最終製品が組織のニーズを満たすかどうかを理解することが含まれる。

このリスクは、AI システムライフサイクルの以下の段階に適用される：

- 入札の募集（AI システムの調達）
- 選定（AI システムの調達）

5.2.5.2 可能な対策

組織はプロバイダに以下の作成を依頼すべきである：

1. 以下の内容を網羅する一般文書：
 - a. AI システムの機能と動作方法：AI システムの動作を詳細に説明する技術仕様書およびアーキテクチャ文書。これには基盤となるアルゴリズム、データ処理方法、機能、既存システムとの統合能力が含まれる。
 - b. ユーザーインターフェースおよびアプリケーションプログラミングインターフェース（API）⁷³の詳細：組織がユーザーおよび開発者の観点から AI システムにアクセスし、利用する方法。
2. AI システムの透明性・解釈可能性・説明可能性に関する文書化：結果が生成される過程が不明瞭な「ブラックボックス」として運用されることを避けるため、組織は AI が結論に至る過程の理解と説明、意思決定を駆動する要因、特定の結果の背景にある明確で理解可能な理由に関する情報を得るべきである。

⁷³異なるソフトウェアアプリケーションがシームレスに通信しデータを交換できるようにするプロトコルとツールの集合体だ。これは仲介役として機能し、様々なソフトウェアコンポーネントが基盤となる実装の詳細を理解する必要なく相互作用することを可能にする。API は、特定の機能やデータにアクセスするための事前定義されたメソッドを提供することで開発プロセスを簡素化する。例えば OpenAI API は、自然言語処理、画像生成、その他の AI タスクのための高度なモデルへのアクセスを提供する。開発者はこの API を利用して、テキスト生成、要約、会話機能などの機能をアプリケーションに統合できる。

3. モデル完全性に関連するサイバーセキュリティ対策に関する文書化：開発過程でモデルの完全性がどのように確保されたか、また継続的な完全性を保証するためにどのような対策が講じられているか／講じるべきか。
4. プロバイダの個人データがバナンス慣行に関する文書化（個人データの収集・処理を含む）：収集したデータの種類は何か？個人データはどのように調達されたか？AI モデルの訓練に個人データがどのように使用されたか、また公平性と正確性を確保するために採用された手法は何か？多くの AI プロバイダーはこの件について曖昧または限定的な情報しか提供しない可能性がある。しかし、訓練用個人データセットの統計的特性を含む、ある程度の透明性は基本である。管理者は、訓練用個人データセットの人口統計的特性が、AI システムが運用中に取り込む個人データの人口統計的特性と近いか遠いかを評価する必要がある。
5. 妥当性確認とテストの手順および結果：モデルは様々なシナリオでどのようにテスト・検証され、その結果はどうだったか？どのデータが使用され、例外ケースはどのように扱われたか？さらに組織は、AI システムを組織の目標に対して評価できる一連の指標をプロバイダに要求すべきだ。

指標は文脈に依存するが、⁷⁴ 一般的な指標には以下がある：

- a. 偽陽性率（FPR）の均等性：誤った陽性事例の数を異なるグループ間で比較できる。この指標は各グループで類似しているべきだ。
- b. 偽陰性率（FNR）の均等性：見逃された真の陽性事例の数を異なるグループ間で比較できる。この指標は各グループで類似しているべきだ。
- c. キャリブレーションの公平性：この指標は、異なるグループ間でモデルの出力と現実を比較する。モデルはグループ間で同等の統計的精度を発揮すべきだ。
- d. 機会均等性（EOP）：類似した特性を持つ個々の入力は、モデルの予測において肯定的な結果を得る機会が同等でなければならない。

これらの一般的な指標に加え、タスク固有のベンチマーク（例：自然言語理解や数学的問題解決）も存在する。これらは AI システムを組織の目標に対して評価することを可能にする。附属書 I には最もよく知られているもののリストが含まれている。

5.3 データ最小化の原則

パターンを正確に学習し、信頼性の高い予測を行い、新規かつ未見のデータに対して良好な汎化を実現するため、AI システムはしばしば大規模なデータセットで訓練される。これは、AI システムがパターンを学習し、受け取った訓練用個人データに近い統計的特性を持つ出力を生成するのに十分な情報を与えるために必要である。もし訓練用個人データが、AI システムが展開後に受け取る入力データを十分に代表していない場合、AI システムは一部の入力データに対して正確な出力を生成できなくなる。

⁷⁴Isabel Barberá, *AI Possible* リスク & 緩和 - *Named Entity Recognition*, September 2023, https://www.edpb.europa.eu/system/files/2024-07/ai-risks_d1named-entity-recognition_edpb-spe-programme_en.pdf Isabel Barberá, *AI Possible* リスク & 緩和 - *Optical Character Recognition*, September 2023, https://www.edpb.europa.eu/system/files/2024-06/ai-risks_d2optical-character-recognition_edpb-spe-programme_en_2.pdf

さらに、訓練用個人データは様々なソースから取得され、世界中に分散し、異なるデータ品質要件を持つ異なる事業体／組織／個人に属している場合がある。EUI は、個人データを用いて AI システムを訓練する前に、有効な法的根拠を確保しなければならない。⁷⁵

EUI はデータ最小化の原則への準拠も確保しなければならない。第 4 条(1)(c)項は、個人データは「処理の目的に関して適切かつ関連性があり、かつ必要最小限に制限されるもの（データ最小化）」と規定している。したがって、AI システムが正確に機能するのに十分な個人データを提供すると同時に、管理者が追求する目的を達成するために必要な量に個人データを制限するというバランスが求められる。

5.3.1 リスク 1 : 個人データの無差別な収集と保存

5.3.1.1 説明

機械学習モデルは訓練用個人データに依存するため、可能な限り多くの訓練用個人データを収集・処理する傾向がある。

明確な規準や関連性なしにあらゆる情報を大量に収集すると、AI システムの目的に不要な情報が蓄積され、データ最小化の原則に反する可能性がある。データ最小化は、関連性があり最新の情報をのみ保存し、意思決定プロセスを歪めたり個人に害を及ぼす可能性のある古いデータや不正確なデータの保持を防ぐ。

このリスクは、AI システムのライフサイクルにおける以下の段階に適用される：

- データ取得と準備（AI システム開発）
- 検証と妥当性確認（AI システムの開発と調達の両方）
- 再評価（AI システムの開発と調達の両方）

5.3.1.2 可能な対策

このリスクに対する可能な対策は以下の通りである：

1. 必要な推論を引き出すのに有用なトレーニング個人データの種類について、事前評価のために当該主題に関する既存情報を利用する。完全なトレーニングと運用前に、計画されたトレーニング個人データタイプの関連性を妥当性確認する。
2. データサンプリング：⁷⁶ トレーニング用個人データの全データセットを使用せず、代表者部分集合をサンプリングする。この手法はデータサンプリングと呼ばれ、データ全体の多様性と主要な特性を正確に反映する、より小さくバランスの取れた部分を選択する。関連する全てのカテゴリーを含み、過剰な代表性やバイアスをかけ

⁷⁵欧州データ保護監察機関、生成的 AI と EUDPR。生成的 AI システム利用時のデータ保護コンプライアンス確保に向けた指針（バージョン 2）、2025 年 10 月 28 日、https://www.edps.europa.eu/system/files/2025-10/2510_28_revised_genai_orientations_en.pdf。

欧州データ保護監督官（EDPS）は既に、ウェブスクレイピング技術を用いた個人データの収集に対して警告を発している。この手法では、個人が自身の知識なく、期待に反し、かつ当初の収集目的とは異なる目的で情報が収集される場合、個人情報に制御不能に陥る恐れがある。EDPS はまた、公開されている個人データの処理も EU データ保護法の対象となることを強調している。この点において、ウェブスクレイピング技術を用いてウェブサイトからデータを収集し、それを訓練目的に使用することは、情報源の信頼性に関する評価がなされていない限り、データ最小化や正確性の原則を含む関連データ保護原則に適合しない可能性がある。」

⁷⁶Daitaku, 「ML モデルのデータダイエット：トレーニングセットの規模が性能に与える影響」, Daitaku ウェブサイト, 2023 年 7 月 19 日, 2025 年 8 月 6 日, <https://blog.dataiku.com/ml-models-on-a-data-diet>

Andrea Montanari, 2024 年 3 月 4 日, 「トレーニングサンプルの最適選択による AI の改善」, Granica.ai ウェブサイト, <https://granica.ai/blog/improving-ai-via-optimal-selection-of-training-samples>

ることなく、慎重にサンプルを設計することで、組織は AI モデルを効果的にトレーニングでき、処理するデータ量を最小限に抑えられる。

3. 匿名化／仮名化：AI システムは可能な限り匿名化されたデータで開発すべきである。個人データが必要な場合は、仮名化されたデータの採用を検討すべきである。

5.4 セキュリティの原則

個人データの安全性を確保する義務は、EUDPR 第 4 条(1)(f)に明記されている：「個人データは、適切な技術的または組織的措置（完全性および機密性）を用いて、不正または違法な処理、ならびに偶発的な喪失、破壊または損傷からの防御を含む、個人データの適切な安全性を確保する方法で処理されなければならない」。

AI コンポーネントを統合した IT システムは、一般的な IT システムに関連するセキュリティ脅威（フィッシング攻撃、マルウェア攻撃など）だけでなく、これらの AI コンポーネントに特有のセキュリティ脅威も考慮しなければならない。

前述の通り、本稿は AI システムの開発・利用に起因するデータ保護上の懸念、リスク、対策に特化して扱う。したがって、対象はデータ保護における AI 固有のセキュリティリスク（一般的な IT システムセキュリティリスクではない）に限定される。⁷⁷

例えば、AI システムの訓練に必要なデータ量を考えると、これらの訓練用個人データは価値が高く、その機密性が侵害された場合、データが漏洩した個人を標的に利用される可能性がある。機密性は、モデル逆転攻撃などの特定の AI 脆弱性を悪用することで侵害される恐れがある。⁷⁸

別の例としては、訓練用個人データ（データ・ポイズニング）や AI モデル自体（モデル・ポイズニング）を操作し、AI システムに誤りを導入することが挙げられる。例えばバイアスを導入したり、AI に無意味な結果を出力させたりする行為だ。⁷⁹ さらに、AI モデル自体が盗まれ、悪意のある目的に利用される可能性もある。

したがって、機密性と完全性（AI システムが意図した通りに機能すること、個人の個人データを保護すること）の観点から、トレーニング用個人データ、入力データ、出力データ、そして AI モデルそのものを防御する必要がある。

5.4.1 リスク 1：AI システムの出力によるトレーニング個人データの開示

5.4.1.1 説明

AI モデルが個人情報を含むデータセットで訓練されると、モデルの出力結果が意図せず訓練セットに含まれる個人の詳細を漏らす可能性がある。この現象は、モデル逆算、メンバーシップ推論攻撃、訓練用個人データの吐き出しといった様々なプライバシー攻撃を通じて発生する。⁸⁰ モデル逆算攻撃では、攻撃者がモデルの出力結果を分析することで機密情報を再構築し、訓練用個人データセット内の個人に関連する個人データを事実上暴

⁷⁷胡裕鵬、鄺文鑫、秦正、李肯利、張吉亮、高彦松、李文佳、李克勤、「人工知能セキュリティ：脅威と対策」、2021 年 11 月 23 日、<https://doi.org/10.1145/3487890>

⁷⁸AI セキュリティ脅威の一種で、攻撃者が機械学習モデルの出力結果を悪用し、その訓練用個人データに関する機密情報を推測する。実質的にモデルをリバースエンジニアリングし、訓練に使用されたデータの機密属性を暴露する手法である。

⁷⁹データ・ポイズニング：敵対者が AI や機械学習モデルが使用する個人データセットを意図的に操作するサイバー攻撃である。虚偽情報の注入、既存データの改変、重要なデータポイントの削除によって、モデルの性能を低下させたり動作を変更したりすることを目的とする。

モデル・ポイズニング：AI モデルの推論段階で特定の悪意ある結果を得るため、敵対者が AI モデルのパラメータやアーキテクチャを意図的に操作する行為。

⁸⁰Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., ... & Xia, S. T., *Privacy leakage on DNNs: A survey of model inversion attacks and defenses*, 11 September 2024, <https://arxiv.org/abs/2402.04013>

露する。メンバーシップ推論攻撃は、AIモデルが生成する信頼度スコアを悪用する。⁸¹ モデルが特定の個人に関する予測に対して高い信頼度を示す場合、攻撃者はその個人がデータセットに含まれていたと推論できる。また、AIモデルが誤って出力例や識別可能な詳細を再現すると、トレーニング個人データからの抜粋や含まれるデータがそのまま吐き出される可能性もある。

このリスクは、AIシステムのライフサイクルにおける以下の段階に適用される：

- 運用と監視（AIシステムの開発と調達の両方）
- 継続的妥当性確認（AIシステムの開発と調達の両方）
- 再評価（AIシステムの開発と調達の両方）

5.4.1.2 可能な対策

このリスクに対する可能な対策は以下の通りである：

1. 個人データの最小化：必要な個人データのみを収集・利用する。これにより、個人を特定できる情報が組み合わされるリスクを最小化する。
2. データ擾乱技術：再識別を困難にするため、トレーニング用個人データを変更する複数の技術が利用可能である。ただし、AIシステムの目的に十分な精度を保持する必要がある：
 - a. 一般化：再識別可能性を減らすため、一部の入力データをより広い範囲で一般化できる。例えば、住所の代わりに郵便番号、市町村名の代わりに県名を使用する。
 - b. 集約：データポイントをグループ化できる。例えば、特定の年齢の代わりに広い年齢層を使用したり、正確な収入の代わりに収入層を使用したりできる。
 - c. 差分プライバシー／ノイズ追加：訓練用個人データに制御されたランダム性を導入する（訓練用個人データの統計的性質は維持する）。
3. 合成データ生成：⁸² AIシステムは、少なくとも部分的に、人工的に生成された個人データを用いて訓練できる。これらの合成データは現実世界のデータの統計的性質を反映しつつ、個人に帰属させられない。⁸³ 適切と判断される場合、この措置は追加的な課題をもたらす可能性があるため、十分な注意を払って実施すべきである。⁸⁴したがって、この措置を検討する場合、追加攻撃（例：メンバーシップ推論攻撃）の可能性を考慮し、上記で提示した追加措置と組み合わせて実施すべきである。⁸⁵

⁸¹Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X., 機械学習に対するメンバーシップ推論攻撃：調査報告, 2022年2月3日, <https://arxiv.org/abs/2103.07853>

⁸²Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., & Wei, W., 機械学習による合成データ生成：レビュー, 2025年4月4日, <https://arxiv.org/abs/2302.04062v9>

⁸³合成データを作成するには、プライバシーと有用性のトレードオフが生じる。このトレードオフを評価する実用的な枠組みは、reslbesl, tandriamil Nampoina Andriamilanto, emidec Emiliano De Cristofaro, bristena-op による「合成データ公開のためのプライバシー評価フレームワーク」で見つかる。2021年6月23日, https://github.com/spring-epfl/synthetic_data_release

⁸⁴Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., ... & Tang, H., 「AIにおける合成データ：課題、応用、倫理的含意」, 2024年1月3日, <https://arxiv.org/abs/2401.01629v1>

⁸⁵Van Breugel, B., Sun, H., Qian, Z., & van der Schaar, M., 「過学習検知による合成データに対するメンバーシップ推論攻撃」, 2023年2月24日, <https://arxiv.org/abs/2302.12580>

4. 出力生成時に、MEMFREE 復号化などの手法を用いて、トレーニング用個人データの完全な複製を防ぐ対策を実施する⁸⁶

5.4.2 リスク 2 : 個人データの保存と個人データ漏えい

5.4.2.1 説明

AI システムに必要な膨大なデータ量はセキュリティリスクを高める。訓練用個人データが何らかの形で侵害された場合（機密性および／または完全性の観点から）、AI システムは深刻な影響を受け、EUDPR 第 3 条(16)の意味におけるデータ侵害を引き起こす可能性がある。⁸⁷ 処理作業全体への影響は、データがどのように影響を受けたかによって異なる。例えば、データの完全性が損なわれた場合（データ・ポイズニングや回避攻撃など）、AI システムは誤動作したり不正確な結果を提供したりする可能性がある。一方、機密性が損なわれた場合、侵害された個人データは個人に影響を与える（侵害された個人データの内容によっては、経済的、健康関連などの影響が生じる可能性がある）。

このリスクは、AI システムのライフサイクルの以下の段階に適用される：

- データ取得と準備（AI システム開発）
- 検証と妥当性確認（AI システムの開発と調達の両方）
- 運用と監視（AI システムの開発と調達の両方）
- 継続的妥当性確認（AI システムの開発と調達の両方）
- 再評価（AI システムの開発と調達の両方）

5.4.2.2 考えられる対策

このリスクに対する可能な対策は以下の通りである：

1. 可能な限り匿名化および／または仮名化を使用する：これにより、機密データ漏洩が発生した場合でも、個人への影響を最小限に抑えられる。ただし、AI システムが効果を発揮するには十分な品質のデータが必要であるという事実とのバランスを取る必要がある。
2. 暗号化：AI システムがデータを積極的に使用していない間はデータを暗号化し、情報漏洩を防ぎ、データの完全性を防御する。
3. 合成トレーニング個人データ：⁸⁸ 実データとは対照的に合成トレーニング個人データを使用することで、機密性の観点から実データが侵害されるリスクを排除できる。合成トレーニング個人データは、開発段階が最終製品の実際の使用に可能な限り合致するよう（出力品質を確保するため）、実データを代表する形で構築されるべきである（例：同一の統計的特性）。この措置は追加的

⁸⁶Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, Nicholas Carlini, 言語モデルにおける逐語的記憶生成の防止は誤ったプライバシー感覚を与える, 2023 年 9 月 11 日, <https://arxiv.org/abs/2210.17546>

⁸⁷つまり、「送信、保存、その他の処理された個人データが、偶発的または違法な破壊、紛失、改ざん、不正開示、または不正アクセスにつながるセキュリティ侵害」である。

⁸⁸Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., & Wei, W., 機械学習による合成データ生成：レビュー, 2025 年 4 月 4 日, <https://arxiv.org/abs/2302.04062v9>

な課題をもたらす可能性があるため、十分な注意を払って実施すべきである。⁸⁹ したがって、この措置を検討する場合は、追加的な攻撃（例：メンバーシップ推論攻撃）の可能性を考慮し、上記で提示した追加措置と組み合わせて実施すべきである。⁹⁰

4. セキュア開発プラクティス：AI モデル開発時にはセキュアコーディングプラクティスに従い、攻撃者がAIコードやインフラの脆弱性を悪用するのを防ぐ。
5. 多要素認証（MFA）：機密性の高いAIシステムへのアクセスにはMFAを導入し、権限のないユーザーによるモデルの改ざんや窃取を防ぐ。

5.4.3 リスク3：アプリケーションプログラミングインターフェース（API）を通じた個人データ漏洩

5.4.3.1 説明

多くのAIシステムは、API呼び出しを通じてアクセス可能なサードパーティプロバイダのAIモデルを使用して構築されている。APIは適切に保護されていない場合、脆弱性がある。これらのAPIへの不正アクセスはデータ漏洩につながる可能性がある。例えば、アクセス制御が適切に定義されていない場合、あるいは詳細なシステム情報を提供するデバッグ用エンドポイントが本番環境で有効なまま放置されている場合、APIが意図せず予定以上のデータを漏洩させる恐れがある。

このリスクは、AIシステムライフサイクルの以下の段階に適用される：

- 運用と監視（AIシステムの開発と調達の両方）

5.4.3.2 可能な対策

このリスクに対する対策としては以下が考えられる：

1. APIへのアクセス：多要素認証（MFA）などの強力な認証メカニズムを導入し、許可されたユーザーとシステムのみがAPIにアクセスできるようにする。
2. 役割ベースのアクセス管理（RBAC）：組織内での役割に基づいて、AIシステムへのアクセス、変更、操作を制限するためにRBACを適用する。
3. スロットリング：⁹¹これは、クライアントがAPIに対して指定された時間枠内で送信できるリクエスト数を制御する技術である。これにより、クライアントの不正利用を防止し、ブルートフォース攻撃などの自動化された攻撃のリスクを緩和する。
4. コミュニケーションの暗号化：クライアントとAPI間で送信されるデータを暗号化するため、HTTPS（TLS）を使用する。これにより、転送中のデータが傍受や盗聴から保護される。
5. ログ記録と監視：API呼び出しのログ記録と監視を実施する。セキュリティ情報イベント管理（SIEM）システムを用いて、潜在的な脅威をほぼリアルタイムで分析し対応する。

⁸⁹Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., ... & Tang, H., AIにおける合成データ：課題、応用、倫理的含意, 2024年1月3日, <https://arxiv.org/abs/2401.01629v1>

⁹⁰Van Breugel, B., Sun, H., Qian, Z., & van der Schaar, M., 「過学習検知による合成データに対するメンバーシップ推論攻撃」, 2023年2月24日, <https://arxiv.org/abs/2302.12580>

⁹¹スロットリングとは、クライアントがAPIに対して一定時間内に送信できるリクエスト数を制御する技術である。これによりトラフィックを効果的に管理し、サーバーの過負荷を防止する。クライアントが許容リクエストレートを超過した場合、スロットリングは一時的にリクエストをブロックまたは遅延させる。これによりリソースの公平な配分が確保され、システム全体のパフォーマンスが維持される。

6. セキュリティ監査：APIの定期的なセキュリティ監査と侵入テストを実施し、脆弱性を識別・対処する。監査にはコードレビュー、設定チェック、脆弱性スキャンを含め、自動化ツールを用いて一般的な脆弱性を継続的にスキャンすべきだ。
7. セキュア開発：入力の妥当性確認やサニタイズなど、API 設計におけるセキュリティのベストプラクティスに従う。
8. パッチ適用：API ソフトウェアと基盤インフラを最新のセキュリティパッチと更新で常に最新の状態に保つ。

5.5 データ対象者の権利

EUDPR はデータ対象者に様々な個別権利をプロバイダしている。具体的には、アクセス権（第 17 条）、訂正権（第 16 条）、消去権（第 19 条）、処理制限権（第 20 条）、データポータビリティ権（第 22 条）、異議申立権（第 23 条）である。AI システムの複雑な性質は、特にデータ対象者がトレーニング用個人データに含まれる個人データに関してこれらの権利を行使する場合、EUI がこれらの権利に対応することをより困難にする可能性がある。モデルがこれらの権利を行使した個人データを「記憶」し、推論段階で再現する可能性があるためである。本節では、データ対象者の権利に関する規定への非遵守に関連する全てのリスクに対処するのではなく、データ管理者がこれらの権利を行使することそのものの可能性を左右する横断的な技術的問題に焦点を当てる。

第一に、これらの権利の実施には処理された個人データの識別が必要だ。例えば、データ対象者が自身の個人データにアクセスできるようにするには、まずシステム内で当該個人データを識別しなければならない。同様に、個人データを消去するにはまず識別が必要だ。管理者がシステム内で個人データを識別して初めて、アクセス提供、修正、消去、制限、ポータビリティを実現できる。第二に、特に訂正権と消去権に焦点を当てると、管理者は識別した個人データを実際に訂正または消去するための措置を実施しなければならない。これは AI システムにおいて特に困難である。なぜなら、トレーニング用個人データがモデルのパラメータに吸収されているためだ。

5.5.1 リスク 1：処理される個人データの不完全な特定

5.5.1.1 説明

AI システムによって個人データが処理される場合、データ対象者は自身の個人データ（トレーニング用個人データセットに含まれ、結果として生成されたモデルに保持されている可能性のあるデータを含む）にアクセスする権利を有する。データ対象者が権利行使を求める要求に応答する際、管理者はトレーニング段階で個人データが使用されているか否か、またその使用箇所を識別するとともに、特定のプロンプトに応答する際にモデルが推論を行う過程で当該データが漏洩する可能性を評価しなければならない。

構造化データセットの場合、⁹² トレーニング個人データが保持されていれば、データ対象者の個人データがトレーニング個人データセット内のどこで処理されているかを識別することは比較的容易である。非構造化データセットの場合、⁹³ 固定されたスキーマやフォーマットが存在しないため、状況はより複雑になる。

AI モデル自体に含まれる個人データを検討する場合、多くの AI モデル（特にデータが複雑な方法で表現・保存される深層学習モデル）の複雑さゆえに、特定のデータポイントへのアクセスは困難である。さらに、同じ要求

⁹²あらかじめ決められた方法で整理され、形式化されたデータである。これにより、コンピューターによる検索、保存、処理が容易になる。IBM, 「構造化データと非構造化データの違いとは?」, IBM ウェブサイト, 2025 年 2 月 7 日, 2025 年 8 月 6 日, https://www.ibm.com/think/topics/structured-vs-unstructured-data?utm_source=chatgpt.com

⁹³あらかじめ定義されたデータモデルを持たない、あるいは行や列のような構造化された方法で組織化されていないデータ。IBM, 「構造化データと非構造化データの違いとは?」, IBM ウェブサイト 2025 年 2 月 7 日, 2025 年 8 月 6 日, https://www.ibm.com/think/topics/structured-vs-unstructured-data?utm_source=chatgpt.com

でも出力が変化し得るといふ本質的特性（AI モデルは本質的に、複数の回答候補から出力を選択する統計的機械である）から、完全な情報の抽出は保証できない。

例えば、ジェーン・ドウが深層ニューラルネットワークで構築されたモデルから自身のデータを削除したい場合、どのモデルパラメータがドウ氏の関連データを表しているかを識別することは不可能である。仮にそれらのパラメータを特定できたとしても、それらはドウ氏と一部の特性を共有する他の個人のデータも表している可能性がある。ドウ氏のデータを消去するためにパラメータを変更することは不可能かもしれない。

このリスクは、AI システムのライフサイクルにおける以下の段階に適用される：

- 開発（AI システムの開発）
- 運用と監視（AI システムの開発と調達の両方）

5.5.1.2 可能な対策

このリスクに対する可能な対策は以下の通りである：⁹⁴

1. 個人データの識別を容易にするためのメタデータの保持：トレーニング用個人データセットには、データ対象者が自身の個人データへのアクセス権を行使した場合に、個人データを含む関連記録やファイルをより容易に識別できるよう、メタデータを含めるべきである。メタデータには、データのソースや収集方法に関する詳細な情報を含めるべきである。さらに、データクリーニング、仮名化、拡張などの前処理手順を文書化し、トレーニング用にデータがどのように準備されたかの明確な記録を確保すべきである。
2. データ取得ツール：データ対象者がアクセス権を行使した際に自身の個人データを提供するため、AI 開発者またはトレーニング用個人データセットプロバイダはデータ取得ツールを作成すべきである。これらのツールは、データ対象者がトレーニング用個人データセット内の自身の個人データについて明確かつ包括的な情報を要求・取得できるようにするものである。当該ツールは、他の利用者の情報を漏洩させることなく、個々のデータレコードを安全に識別・取得できるように設計されなければならない。これらのツールが提供する個人データは、機械可読かつユーザーフレンドリーな形式であるべきだ。
3. MemHunter（LLM の記憶検知自動ツール）のようなツールを導入できる。⁹⁵

5.5.2 リスク 2：不完全な訂正または消去

5.5.2.1 説明

AI システムによって個人データが処理される場合、データ対象者は、AI システムによって処理された誤った個人データの訂正、および／または AI システムからの個人データの消去を要求する権利を有する。データ対象者が、AI システムが自身に関する誤ったデータまたは不完全なデータを保有していると考えた場合、それが AI システムのトレーニング用個人データセット内にあるか、AI システムの出力結果内にあるかを問わず、組織に対してその修正を要求できる。

これらの二つの権利の行使は、アクセス権の行使と同様の困難を伴う。データがデータセットやモデル内で識別できない場合、その消去や訂正は不可能である。構造化データセットの場合、トレーニング用個人データが保持さ

⁹⁴クリス・シュリヤク博士、「AI：複雑なアルゴリズムと効果的なデータ保護監督」、EDPB ウェブサイト、2024 年 3 月、https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-complex-algorithms-andeffective-data_en

⁹⁵Zhenpeng Wu, Jian Lou, Zibin Zheng, Chuan Chen, 「MemHunter：LLM におけるデータセット規模での自動化・検証可能な記憶検知」, 2024 年 12 月 10 日、<https://arxiv.org/html/2412.07261v1>

れていれば（データの構造化特性ゆえに）、データ対象者の個人データを訂正または消去することは比較的容易である。一方、非構造化データセットでは、固定されたスキーマやフォーマットが存在しないため、状況はより複雑になる。

AI システムの出力結果を修正することは、特にデータが深層ニューラルネットワークや LLM のような複雑なモデルに組み込まれている場合、課題となる。消去権も AI システムに対して同様の複雑さを生じさせる。AI モデルの性質上、これらの訂正・消去要求を AI モデルに完全に実装することは複雑になり得る。AI モデルは膨大なデータセットから学習することが多く、一度訓練されると、分離・訂正・消去が困難なパターンや情報を保持し続ける可能性がある。

このリスクは、AI システムライフサイクルの以下の段階に適用される：

- 開発（AI システムの開発）
- 運用・監視（AI システムの開発と調達の両方）

5.5.2.2 可能な対策

このリスクに対する可能な対策は以下の通りである：

1. データ取得ツール：5.6.1.2 項と同様に、開発者または供給者は、トレーニング用個人データセット内の個人データを修正または消去できるように、データ修正・消去ツールを作成すべきである。
2. マシン・アンラーニング：⁹⁶ マシン・アンラーニングとは、機械学習モデルが以前に学習した特定のデータポイントを選択的に忘却するプロセスであり、モデルがそのデータで訓練されたことがなかったかのように振る舞うことを可能にする。マシン・アンラーニングは主に二つの手法に分類される。完全なアンラーニングは、指定されたデータの影響を完全に除去するためモデルをゼロから再学習させる手法である。近似アンラーニングは、モデルのパラメータを限定的に更新することでデータの影響を最小化しようとする手法である。完全なアンラーニングはデータ除去の強力な保証を提供する一方、必要な変更により計算上または金銭的に高コストとなることが多い。一方、近似アンラーニングはより効率的な代替手段となるが、データの影響を完全に除去できない可能性がある。
3. 機械学習によるアンラーニングが不可能な場合、出力フィルタリングが用いられる。出力フィルタリングは、AI モデルの応答をリアルタイムでスキャンし、個人を特定できる情報に到達する前に検知・遮断する手法である。システムはパターン認識や事業体検出を活用し、個人データを識別してユーザーに届く前にブロックする。

⁹⁶EDPS, *Techsonar 2025*, 2024 年 11 月 15 日, https://www.edps.europa.eu/data-protection/ourwork/publications/reports/2024-11-15-techsonar-report-2025_en

6 結論

パーソナルデータの処理を行う AI システムを運用することは、データ対象者にとって重大なリスクを伴う。EUI は管理者として、これらのリスクを識別・評価・緩和する法的・倫理的義務を負っている。その影響は甚大である：適切な安全対策が当初から組み込まれていない場合、AI システムは前例のない規模と速度で基本的人権へのリスクを増幅させる可能性がある。このため、本稿では ISO 31000:2018 に準拠したリスクマネジメント手法を適用し、EUDPR の特定要件に合わせて文脈化することで、リスクが体系的かつ説明責任のある方法で対処されるよう支援した。

第 2 章では、データ保護課題に対処する基盤としてリスクマネジメントを紹介し、脅威を分析し適切な安全対策を実装するための構造化された基盤を提供した。第 3 章では AI システムの定義とライフサイクルを検討し、システム運用前にリスクを予測できる決定的な段階として調達を強調した。第 4 章では、透明性、公正性、正確性、データ最小化、セキュリティという 5 つのデータ保護原則を検証し、これらの原則への準拠を確保するためにリスクがどのように顕在化するかを分析した。これには、データ対象者の権利の効果的な行使に関連するリスクも含まれる。各原則について、文書は具体的なリスクシナリオを識別し、非網羅的な技術的対策例を示して、実践的なリスクマネジメント手法を説明した。

本分析は、あらゆるリスクと緩和策を網羅的に提示することを意図していない。むしろ、EU 域内事業者がリスクマネジメントへの体系的なアプローチを構築する実践的枠組みを提供し、他のリスクやコンプライアンス要件を考慮して補完する必要性を示唆するものである。したがって、AI システムが運用される多様な状況に適応可能な枠組みとして設計されている。最終的には、管理者自身が包括的なコンプライアンス評価を実施する全責任を負い、アセスメントを実施し、必要な法的分析で本枠組みを補完し、自らの決定が EUDPR と整合することを確保する責任を負う。

結論として、AI システムが提起するリスクへの対応は、EUI にとって周辺的な課題ではなく中核的な義務である。EUDPR への準拠、基本的権利の防御、公衆の信頼維持は全て、管理者が AI ライフサイクル全体を通じてリスクを積極的に識別・評価・緩和する能力にかかっている。これは単発の評価では不十分であり、説明責任の文化、継続的な監視、適応的な改善を求めるものである。こうした実践を AI ガバナンスに組み込むことで、EUI は AI 開発の複雑性を乗り越え責任ある AI 利用を実現できるだけでなく、イノベーションが基本権とデータ保護原則の尊重に確固として根ざすよう確保するリーダーシップを示すことができる。

附属書 1 : 評価指標

AI 性能評価の指標は、AI システムの種別によって大きく異なる。これは AI アプリケーションの多様性と各々の目的を反映している。⁹⁷

例えば、自然言語処理タスクでは、生成されたテキストや翻訳の品質を評価するために、BLEU、ROUGE、GLEU、METEOR といった指標が一般的に用いられる。これらの指標は、AI が生成した出力を参照テキストと比較することに焦点を当て、精度、再現率、意味的類似性といった側面を測定する。

一方、AI モデルの分類や識別タスクでは、統計的精度、精度、再現率、F1 スコアといった指標がよく用いられる。これらの指標は、AI モデルがデータを事前定義されたクラスや個体にどれだけ適切に分類できるかを評価するもので、スパム検知や画像認識などの応用において重要だ。

AI が連続値を予測する回帰問題では、平均絶対誤差 (MAE)、平均二乗誤差 (MSE)、二乗平均平方根誤差 (RMSE) といった異なる指標が用いられる。

ベンチマークは、異なる AI モデルを評価・比較するための標準化された手法を提供する。⁹⁸ 一貫したデータセット、事前定義されたタスク、評価指標を提供することで、研究者や開発者は AI ソリューションの性能、効率性、統計的精度を客観的に評価できる。さらにベンチマークは、AI システムの使用ライフサイクル全体で対処すべき懸念領域を特定することで、導入前の AI モデルの潜在的なリスクや限界を明らかにする上で重要な役割を果たす。

以下の表は、本報告書の公開時点で利用可能な AI システム評価用ベンチマークの一部を列挙したものである。このリストは網羅的なものではないが、良い出発点となる。

AI の種類	利用可能なベンチマーク	概要
自然言語処理 (NLP) と大規模言語モデル (LLM) ⁹⁹	リコール重視の要約評価指標 (ROUGE) ¹⁰⁰	自然言語処理における自動生成要約と機械翻訳の品質を評価するために設計された指標群とソフトウェアパッケージ。
	バイリンガル評価	機械生成テキストの品質を評価するベンチマーク。特に機械翻訳において用いられる。
	Understudy (BLEU) ¹⁰¹	翻訳タスクにおいて機械生成テキストの品質を評価するベンチマークである。生成テキストと参照テキストの両方に現れる n-gram (連続する n 個の単語の列) の精度を測定することで、この類似度を算出する。
	General Language Understanding	様々な言語理解タスクにおける NLP モデルの性能を評価するために設計されたベンチマークデータセットである。最初に導

⁹⁷ OECD.ai、「信頼できる AI のためのツールと指標のカタログ」、2025 年 8 月 6 日、[OECD.ai ウェブサイト、
https://oecd.ai/en/catalogue/metrics](https://oecd.ai/en/catalogue/metrics)

⁹⁸ Github、「コード付き論文」、Github ウェブサイト、2025 年 8 月 6 日、<https://paperswithcode.com/sota>

⁹⁹ Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, Xing Xie. 2024 年。大規模言語モデルの評価に関する調査、2024 年 3 月 29 日、<https://doi.org/10.1145/3641289>

¹⁰⁰ ネリ・ヴァン・オッテン、「NLP における ROUGE 指標：完全ガイドと Python での実践チュートリアル」、2024 年 8 月 12 日、2025 年 8 月 6 日、<https://spotintelligence.com/2024/08/12/rouge-metric-in-nlp/>

¹⁰¹ <https://spotintelligence.com/2024/08/13/bleu-score-in-nlp/>

AIの種類	利用可能なベンチマーク	概要
	(GLUE) ¹⁰² と SuperGLUE ¹⁰³	入された GLUE は、文分類、感情分析、テキスト含意など、9つの多様な NLP タスクで構成されている。GLUE の後継として開発された SuperGLUE は、より高度な推論、常識知識、文脈理解を必要とする複雑なタスクを導入し、前身を発展させたものである。GLUE が比較的単純な言語課題に焦点を当てる一方、SuperGLUE は質問応答、共参照解決 ¹⁰⁴ 、読解などのタスクを取り入れ、モデルに高次認知能力の証明を求める。
	言語モデルの包括的評価 (HELM) ¹⁰⁵	幅広いシナリオと指標で言語モデルの能力、限界、潜在的なリスクを評価するために設計されたベンチマーク枠組みだ。この枠組みは 16 のコアシナリオと 26 のターゲットシナリオでモデルを評価し、統計的精度、キャリブレーション、頑健性、公平性、バイアス、有害性、効率性の 7 つの主要指標を測定する。
	大規模マルチタスク理解 (MMLU) ¹⁰⁶ および MMLU-Pro ¹⁰⁷	MMLU は約 16,000 の多肢選択式問題で構成され、数学、哲学、法律、医学を含む 57 の学術分野を網羅する。このベンチマークは、初級から専門家レベルまでの難易度で、AI モデルの一般知識と問題解決能力を評価することを目的としている。 MMLU-Pro はより難易度の高い推論重視の問題を特徴とし、選択肢を 4 つから 10 つに増やしている。
画像認識 ¹⁰⁸	ImageNet ¹⁰⁹	視覚的物体認識研究用に設計された視覚データベースである。数千の物体カテゴリーを網羅する 1400 万枚以上のラベル付き画像を含む。データセットは 1000 の異なるクラスに構

¹⁰²Wang, A., *Glue: A multi-task benchmark and analysis platform for natural language understanding*, 2019 年 2 月 22 日, <https://arxiv.org/abs/1804.07461>

¹⁰³Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S., *Superglue: 汎用言語理解システムのためのより粘着性の高いベンチマーク*. 神経情報処理システムの進歩, 2020 年 2 月 13 日, <https://arxiv.org/abs/1905.00537>

¹⁰⁴同一の事業体や事象を指す表現をテキスト内で識別し関連付けること

¹⁰⁵Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y., *言語モデルの包括的評価*, 2023 年 10 月 1 日, <https://arxiv.org/abs/2211.09110>

¹⁰⁶Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J., *Measuring massive multitask language understanding*, 2021 年 1 月 12 日, <https://arxiv.org/abs/2009.03300>

¹⁰⁷Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., ... & Chen, W., *Mmlu-pro: A more robust and challenging multitask language understanding benchmark*, 2024 年 11 月 6 日, <https://arxiv.org/abs/2406.01574v4>

¹⁰⁸Li, L., Chen, G., Shi, H., Xiao, J., & Chen, L. (2024). *マルチモーダルベンチマークに関する調査：大規模 AI モデルの時代において*, 2024 年 9 月 21 日, <https://arxiv.org/abs/2409.18142>

Rangel, Gabriela, Cuevas-Tello, Juan C., Nunez-Varela, Jose, Puente, Cesar, Silva-Trujillo, Alejandra G., *A Survey on 畳み込みニューラルネットワーク and Their Performance Limitations in Image Recognition Tasks*, 2024 年 7 月 12 日, <https://doi.org/10.1155/2024/2797320>

¹⁰⁹Image Net. 「Imagenet Database」, Image Net ウェブサイト, 2021 年 3 月 11 日, 2025 年 8 月 6 日, <https://www.image-net.org/>

AIの種類	利用可能なベンチマーク	概要
		造化されており、約 120 万枚の画像が訓練用、5 万枚が妥当性確認用、10 万枚がテスト用として使用される。
	CIFAR-10 および CIFAR-100 ¹¹⁰	<p>CIFAR-10 は、10 の互いに排他的なクラスに分類された 6 万枚の 32x32 カラー画像で構成される。データセットは 5 万枚の訓練画像と 1 万枚のテスト画像に分割され、各クラスは均等に代表されている。</p> <p>CIFAR-100 は、CIFAR-10 と同じ総画像数と画像サイズを維持しつつ、分類課題を拡大するため、データを 100 クラスに分割している。これらのクラスはさらに 20 のスーパークラスにグループ化され、追加の分類層を提供している。CIFAR-100 の各画像は、「細かいラベル」（特定のクラス）と「粗いラベル」（上位クラス）の両方に関連付けられている。</p>
	MNIST ¹¹¹	このデータセットは、手書き数字のグレースケール画像 70,000 枚で構成され、各画像は 28x28 ピクセルのサイズである。これは 2 つのサブセットに分けられている：60,000 枚の画像からなるトレーニングセットと、10,000 枚の画像からなるテストセットである。

カテゴリー	ベンチマーク	説明	テスト対象モデル
自然言語処理 (NLP)	GLUE (一般言語理解評価)	モデルの一般的な言語理解能力をテストするために設計されたタスクの集合体だ。	テキストベースのモデル、言語モデル (例: BERT、GPT)
	SuperGLUE	より難易度の高いタスクを追加した GLUE の拡張版だ。	高度な NLP モデル (例: T5、RoBERTa)
	SQuAD (スタンフォード質問応答データセット)	文章に基づく読解力と質問への回答能力をテストする。	QA モデル (例: BERT、T5)
	CoNLL-03 ¹¹²	事業体認識 (NER) 性能評価用データセット。	NER モデル (例: LSTM、CRF)

¹¹⁰Alex Krizhevsky, 「CIFAR-10 データセット」, Alex Krizhevsky のホームページ, 2025 年 8 月 6 日, <https://www.cs.toronto.edu/~kriz/cifar.html>

¹¹¹Hojjat Khodabakhsh, 「MNIST Dataset」, Kaggle ウェブサイト, 2025 年 8 月 6 日, <https://www.kaggle.com/datasets/hojjatk/mnist-dataset>

¹¹²Github, 「コード付き論文」, Github ウェブサイト, 2025 年 8 月 6 日, <https://paperswithcode.com/cobll-2023>

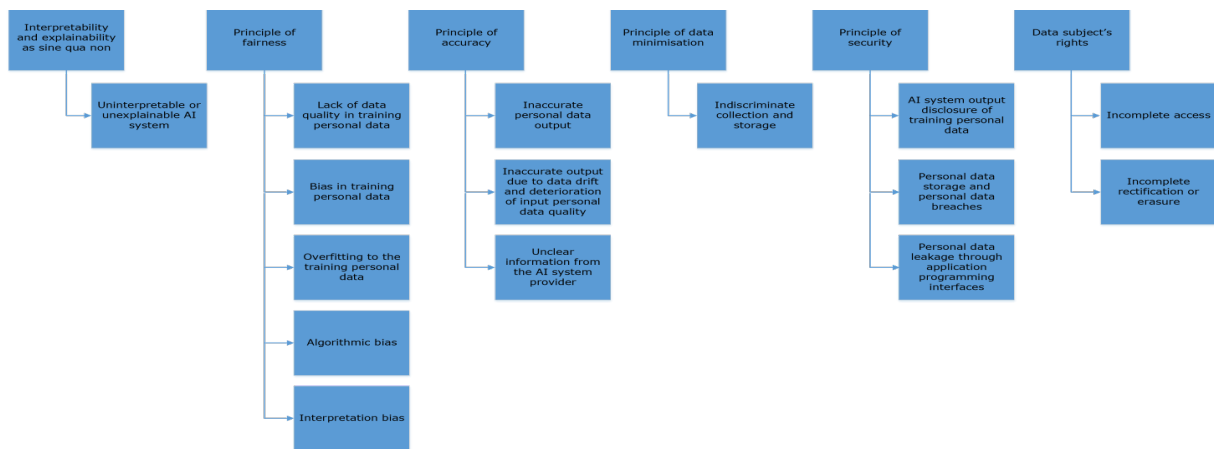
カテゴリー	ベンチマーク	説明	テスト対象モデル
	MNLI (多ジャンル自然言語推論)	モデルが前提が仮説を包含するか、矛盾するか、あるいは中立かを判断する能力を評価する。	推論モデル (例: BERT、RoBERTa、XLNet)
	TREC (テキスト検索会議)	質問分類タスクに焦点を当てる。	テキスト分類器、意図認識モデル
コンピュータビジョン (CV)	ImageNet	膨大な数のカテゴリー (1000) を持つ大規模画像分類ベンチマーク。	CNN、ビジョントランスフォーマー (例: ResNet、EfficientNet)
	COCO (Common Objects in Context) ¹¹³	物体検知、セグメンテーション、キャプション付けタスクのベンチマークである。	物体検知およびセグメンテーションモデル (例: YOLO、Mask R-CNN、Faster R-CNN)
	PASCAL VOC	画像分類、物体検知、セグメンテーションタスクに焦点を当てている。	物体検知、セグメンテーションモデル
	ADE20K	高密度なピクセル単位の注釈を扱うセマンティックセグメンテーションのベンチマークデータセットである。	セグメンテーションモデル (例: DeepLabV3+、U-Net)
	KITTI	自動運転に焦点を当てたデータセット。ステレオマッチング、オプティカルフロー、物体検知などのタスクを含む。	物体検知、オプティカルフローモデル (例: CNN、RNN)
音声とオーディオ	LibriSpeech	英語のオーディオブックの文字起こしに特化した音声認識ベンチマーク。	音声認識モデル (例: DeepSpeech、Wav2Vec)
	VoxCeleb	音声クリップにおける話者認識と識別。	話者認識モデル (例: ECAPA-TDNN、VGGVox)
	TIMIT	音響音声学 連続音声認識のためのコーパス。	音声認識モデル

¹¹³Cocodataset, 「Cocodataset」, Cocodataset ウェブサイト, 2025 年 8 月 6 日, <https://cocodataset.org>

カテゴリー	ベンチマーク	説明	テスト対象モデル
	CHiME	騒がしい環境での音声認識を評価する。	頑健な音声認識モデル (例：RNN、LSTM)
強化学習	OpenAI Gym	様々な環境で強化学習アルゴリズムを開発・比較するためのプラットフォームだ。	強化学習エージェント (例：PPO、DDPG、A2C)
	MuJoCo	連続制御タスクの環境シミュレーションに用いられる物理エンジン。	連続制御用強化学習エージェント (例：DDPG、TRPO)
	視覚的質問応答 (VQA)	画像に関する自然言語の質問に答えるモデルの能力をテストする。	ビジョン+NLP モデル (例：LXMERT、ViLBERT)
マルチモーダル AI	MS COCO キャプション生成	画像の自然言語による説明文生成を評価する。	画像キャプション生成モデル (例：Show and Tell、Image Transformer)
	AI2 推論チャレンジ (ARC)	幅広いトピックに関する多肢選択問題で、一般的な推論能力をテストするベンチマークだ。	汎用推論 AI (例：GPT、T5)
汎用 AI の性能	CLIP (対照的言語-画像事前学習)	モデルがテキストと画像の入力をどれだけ適切に一致させられるかを測定する。ゼロショット分類などのタスクに適用される。	マルチモーダル AI モデル (例：CLIP、Flamingo)
	WinoBias	言語モデルが性別代名詞にどう反応するかを分析し、バイアスを測定する。	NLP モデル (例：GPT、BERT)
公平性とバイアス	FairFace	顔認識モデルが、異なる人口統計学的特性 (肌の色、年齢、性別など) に対して公平かどうかを評価する。	顔認識モデル (例：FaceNet、ArcFace)
	MedNLI	医療分野における自然言語推論を評価するためのベンチマーク。	医療 NLP モデル (例：BioBERT、ClinicalBERT)

カテゴリー	ベンチマーク	説明	テスト対象モデル
医療向け AI	ChestX-ray14	X 線画像から 14 種類の一般的な胸部疾患を検知するモデルの評価に使用される。	医療画像分類モデル (例：CNN)
	MIMIC-CXR	診断統計的精度を評価するための大規模胸部 X 線データセット。	医療画像モデル (例：ResNet、DenseNet)

附属書 2 : 懸念事項とリスクの概要



解釈可能性と説明可能性は不可欠な条件	解釈不能または説明不能な AI システム
公平性の原則	訓練用個人データの品質不足
	訓練用個人データにおけるバイアス
	訓練用個人データへの過学習
	アルゴリズムバイアス
	解釈バイアス
正確性の原則	不正確な個人データ出力
	データドリフト及び入力個人データ品質の劣化による不正確な出力
	AI システムプロバイダからの不明確な情報
データ最小化の原則	無差別な収集及び保存
セキュリティの原則	AI システム出力訓練用個人データの開示
	個人データの保存及び個人データ漏えい
	アプリケーションプログラミングインターフェースを通じた個人データ漏洩
データ対象者の権利	不完全なアクセス
	不完全な修正または消去

附属書 3 : AI ライフサイクル開発の各段階におけるチェックリスト

AI システムの開発

AI ライフサイクル開発の段階	原則	リスク
1. 構想／分析	5.1 公平性の原則	5.1.4 アルゴリズムバイアス
2. データ取得と準備	5.1 公平性の原則	5.1.1 個人データ訓練におけるデータ品質の欠如
		5.1.2 トレーニング用個人データにおけるバイアス
		5.1.3 トレーニング個人データへの過学習
	5.2 正確性の原則	5.2.3 不正確な個人データの出力
	5.3 データ最小化の原則	5.3.1 無差別な収集と保管
	5.4 セキュリティの原則	5.4.2 個人データの保管と個人データ漏えい
3. 開発	5.5 データ対象者の権利	5.5.1 不完全なアクセス
		5.5.2 不完全な訂正または消去
4. 検証と妥当性確認	4 解釈可能性と説明可能性	4.1 解釈不能または説明不能な AI システム
	5.1 公平性の原則	5.1.1 個人データの訓練におけるデータ品質の欠如
		5.1.2 トレーニング用個人データにおけるバイアス
		5.1.3 トレーニング個人データへの過学習
		5.1.4 アルゴリズムバイアス
		5.1.5 解釈バイアス
	5.2 正確性の原則	5.2.3 不正確な個人データの出力
	5.3 データ最小化の原則	5.3.1 無差別な収集と保管
5.4 セキュリティの原則	5.4.2 個人データの保管と個人データ漏えい	
5. 展開	なし	なし
6. 運用と監視	4 解釈可能性と説明可能性	4.1 解釈不能または説明不能な AI システム

	5.1 公平性の原則	5.1.3 訓練用個人データへの過学習
		5.1.5 解釈バイアス
	5.2 正確性の原則	5.2.3 不正確な個人データの出力
		5.2.4 データドリフト及び入力個人データの品質劣化による不正確な出力
	5.4 セキュリティの原則	5.4.1 訓練用個人データの AI システム出力開示
		5.4.2 個人データの保管と個人データ漏えい
		5.4.3 アプリケーションプログラミングインターフェースを通じた個人データの漏洩
	5.5 データ対象者の権利	5.5.1 不完全なアクセス
		5.5.2 不完全な訂正または消去
	7. 継続的妥当性確認	4 解釈可能性と説明可能性
5.1 公平性の原則		5.1.3 訓練用個人データへの過学習
		5.1.4 アルゴリズムバイアス
5.2 正確性の原則		5.2.4 データドリフト及び入力個人データの品質劣化による不正確な出力
5.4 セキュリティの原則		5.4.1 トレーニング個人データの AI システム出力開示
	5.4.2 個人データの保管と個人データ漏えい	
8. 再評価	4 解釈可能性と説明可能性	4.1 解釈不能または説明不能な AI システム
	5.1 公平性の原則	5.1.1 個人データ訓練におけるデータ品質の欠如
		5.1.2 トレーニング用個人データにおけるバイアス
		5.1.3 トレーニング個人データへの過学習
		5.1.4 アルゴリズムバイアス
		5.1.5 解釈バイアス
5.2 正確性の原則	5.2.3 不正確な個人データの出力	

		5.2.4 データドリフト及び入力個人データの品質劣化による不正確な出力
	5.3 データ最小化の原則	5.3.1 無差別な収集と保管
	5.4 セキュリティの原則	5.4.1 トレーニング用個人データの AI システム出力開示
		5.4.2 個人データの保管と個人データ漏えい
9. 廃止	なし	なし

AI システムの調達

AI ライフサイクル開発の段階	懸念事項	リスク
1. 準備	なし	なし
2. 入札の募集	5.2 正確性の原則	5.2.5 AIシステムプロバイダからの不明確な情報
3. 選定	4 解釈可能性と説明可能性	4.1 解釈不能または説明不能な AI システム
	5.2 正確性の原則	AI システムプロバイダからの不明確な情報
4. 授与と契約	なし	なし
5. 履行	なし	なし
6 検証と妥当性確認	4 解釈可能性と説明可能性	4.1 解釈不能または説明不能な AI システム
	5.1 公平性の原則	5.1.1 個人データの訓練におけるデータ品質の欠如
		5.1.2 トレーニング用個人データにおけるバイアス
		5.1.3 トレーニング個人データへの過学習
		5.1.4 アルゴリズムバイアス
		5.1.5 解釈バイアス
	5.2 正確性の原則	5.2.3 不正確な個人データの出力
5.3 データ最小化の原則	5.3.1 無差別な収集と保管	
5.4 セキュリティの原則	5.4.2 個人データの保管と個人データ漏えい	

7 展開	なし	なし
8 運用と監視	4 解釈可能性と説明可能性	4.1 解釈不能または説明不能な AI システム
	5.1 公平性の原則	5.1.3 訓練用個人データへの過学習
		5.1.5 解釈バイアス
	5.2 正確性の原則	5.2.3 不正確な個人データの出力
		5.2.4 データドリフト及び入力個人データの品質劣化による不正確な出力
	5.4 セキュリティの原則	5.4.1 トレーニング個人データの AI システム出力開示
		5.4.2 個人データの保管と個人データ漏えい
		5.4.3 アプリケーションプログラミングインターフェースを通じた個人データの漏洩
	5.5 データ対象者の権利	5.5.1 不完全なアクセス
		5.5.2 不完全な訂正または消去
9 継続的妥当性確認	4 解釈可能性と説明可能性	4.1 解釈不能または説明不能な AI システム
	5.1 公平性の原則	5.1.3 訓練用個人データへの過学習
		5.1.4 アルゴリズムバイアス
		5.2.4 データドリフト及び入力個人データの品質劣化による不正確な出力
	5.4 セキュリティの原則	5.4.1 訓練用個人データの AI システム出力開示
		5.4.2 個人データの保管と個人データ漏えい
	10 再評価	4 解釈可能性と説明可能性
5.1 公平性の原則		5.1.1 個人データ訓練におけるデータ品質の欠如
		5.1.2 トレーニング用個人データにおけるバイアス
		5.1.3 トレーニング個人データへの過学習
		5.1.4 アルゴリズムバイアス

		5.1.5 解釈バイアス
	5.2 正確性の原則	5.2.3 不正確な個人データの出力
		5.2.4 データドリフト及び入力個人データの品質劣化による不正確な出力
	5.3 データ最小化の原則	5.3.1 無差別な収集と保管
	5.4 セキュリティの原則	5.4.1 トレーニング用個人データの AI システム出力開示
		5.4.2 個人データの保管と個人データ漏えい
11 廃止	なし	なし