EU・英国・米国におけるAI倫理動向

2020年11月4日

(株)国際社会経済研究所　調査研究部
主幹研究員　小泉 雄介

Institute for International Socio-Economic Studies 2020

目 次

○ EU
- 欧州委員会のAI倫理に向けた取り組み
- AI倫理ガイドライン／信頼できるAIのためのアセスメントリスト
- AIホワイトペーパーのリモート生体識別（顔認識等）に対するEDPS見解
- 米国顔認識ベンダーClearview AIに対するEDPB意見

○ 英国（イギリス）
- 英国のAI倫理・データ保護関連の報告書・ガイダンス類
- ICOのAIとデータ保護に関するガイダンス

○ 米国（アメリカ）
- NISTのAIワークショップシリーズ
- NISTの「説明可能なAIの4つの原則」ドラフト

（参考資料）
- 欧州委員会のAIホワイトペーパー 概要
- 英国ICOのAIとデータ保護に関するガイダンス 概要

欧州委員会のAI倫理に向けた取り組み

- 2018年3月9日： 欧州委員会の科学新技術倫理グループ（EGE）が「AI・ロボティクス・自律システムに関する宣言」を公表。
- 2018年4月25日： 欧州委員会はAI戦略「欧州のAI」を公表。
- 2019年4月8日： 欧州委員会の設立したAIハイレベル専門家グループが「信頼できるAIのための倫理ガイドライン」を公表。以下の7つの要件を規定。
  - 人間の主体性と監督
  - 技術的な堅牢性と安全性
  - プライバシーとデータガバナンス
  - 透明性
  - 多様性、非差別、公平性
  - 社会および環境の福祉
  - 説明責任（アカウンタビリティ）
- 2019年後半： 上記AI倫理ガイドラインのアセスメントリストについて、350以上の組織がテストを行い、フィードバックを回答。
- 2020年7月17日： AIハイレベル専門家グループはフィードバックに基づき、アセスメントリストの改訂版を公表。
- 2020年2月19日： 欧州委員会は「AIに関するホワイトペーパー：優越と信頼に向けた欧州アプローチ」を公表。
- 2020年7月23日： 欧州委員会はAIホワイトペーパーを補完する「Inception Impact Assessment」を公表。
- 2020年10月20日： 欧州議会は「AI・ロボティクス・関連技術の倫理フレームワークに関する欧州委員会への勧告」を決議。
- 2021年１Ｑ： 欧州委員会がAIホワイトペーパーに対する法規制面でのフォローアップ文書を公表予定

信頼できるAIのためのアセスメントリスト

- **欧州委員会のAIハイレベル専門家グループは2020年7月17日、信頼できるAIのためのアセスメントリストの改訂版を公表。**
  (https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment)
- **対象者はAI開発企業およびAIディプロイヤー。アセスメントリストにより、AI倫理ガイドラインで示されたAI原則や7要件を実装に落とし込むことが可能に。**

7つの要件	サブカテゴリ	概要
1. 人間の主体性と監督	・人間の主体性と自律 ・人間による監督	AIシステムは、「人間の自律の尊重の原則」で規定したように、人間の自律と意思決定をサポートするべきである。このためには、AIシステムは、利用者の主体性(agency)をサポートし、基本的人権を促進することによって民主的で繁栄的で平等な社会を実現させるものとして働くべきであり、人間による監督を可能とするものであるべきである。
2. 技術的な堅牢性と安全性	・攻撃からの回復力とセキュリティ ・全般的安全性 ・正確性 ・確実性、予備プラン、再現性	信頼できるAIを実現するための極めて重要な要素は、技術的な堅牢性である。それは「害悪防止の原則」と緊密に結びついている。技術的な堅牢性は、AIシステムがリスク予防的アプローチで開発され、意図しない・予期しない害悪を最小化し、受け入れがたい害悪を防止しながら、意図したように確実に動作する仕方で開発されることを求めるものである。このことは、運用環境の可能な変化や、AIシステムと敵対的なやり取りをする(人間および人工の)エージェントの存在にも対応できるべきである。さらに、人間の物理的・精神的な完全性(integrity)が保証されるべきである。
3. プライバシーとデータガバナンス	・プライバシー ・データガバナンス	プライバシーという基本的人権は、とりわけAIシステムによって影響されやすい。「害悪防止の原則」と緊密に結びついている。プライバシーへの害悪を防止するためには、以下に対する十分なデータガバナンスを必要とする。すなわち、利用するデータの品質と一貫性、AIシステムが配置される領域におけるデータの関連性、AIシステムのアクセスプロトコル、データを処理する能力に対するガバナンスである。

信頼できるAIのためのアセスメントリスト

7つの要件	サブカテゴリ	概要
4. 透明性	・トレーサビリティ	透明性の要件は「説明可能性の原則」と緊密に結びついている。データの透明性、システムの透明性、ビジネスモデルの透明性を包含する。
	・説明可能性（explainability）	
	・コミュニケーション	
5. 多様性、非差別、公平性	・不公平なバイアスの回避	信頼できるAIを実現するために、AIシステムのライフサイクル全体を通じて包摂性と多様性を実現しなければならない。プロセスを通じて影響を受ける全てのステークホルダーの考慮と関与に加えて、設計プロセスへの平等なアクセスと、平等な取扱いを保証するべきである。この要件は「公平性の原則」と緊密に結びついている。
	・アクセシビリティとユニバーサルデザイン	
	・ステークホルダーの参加	
6. 社会および環境の福祉	・環境へのやさしさ	「公平性の原則」および「害悪防止の原則」に則り、社会全体や、他の感覚ある生物、環境もまた、AIシステムのライフサイクル全体を通じたステークホルダーとしてみなされるべきである。AIシステムの持続可能性や環境的責任が促進されるべきであり、SDGsのようなグローバルな課題に対処するAIソリューションの研究が促進されるべきである。理想的には、AIシステムは将来の世代も含め、全ての人間にベネフィットを与えるために使われるべきである。
	・仕事とスキルへの影響	
	・社会全般への影響と民主主義	
7. 説明責任（アカウンタビリティ）	・監査可能性	説明責任の要件は、上記の6要件を補完するものである。これは、「公平性の原則」と緊密に結びついている。AIシステムの開発・配置・利用の前と後において、AIシステムとその成果物における責任（responsibility）と説明責任（accountability）を保証するための適切なメカニズムが必要である。
	・リスクマネジメント	

信頼できるAIのためのアセスメントリスト（項目例）

○要件1．人間の主体性と監督

◇ 人間の主体性と自律

- このサブセクションでは、AIシステムが最も広い意味で人間の行動に与えうる影響について扱う。すなわち、意思決定プロセスにおいて人間のガイドやサポートを目的としたAIシステムの影響を扱う。アルゴリズムによる意思決定支援システムやリスク分析/予測システムを含む（レコメンドシステム、予測的警察活動、財務リスク分析など）。また、人間のように「行為」するAIシステムに直面したときの人間の知覚や期待への影響も扱う。最後に、AIシステムが人間の愛情、信頼、および（非）依存に及ぼす影響を扱う。

- AIシステムは、人間や社会に影響を与える意思決定について、人間のエンドユーザとの相互作用や、エンドユーザのガイドや、エンドユーザによる意思決定を行うように設計されていますか？
  - AIシステムは、意思決定やコンテンツ・アドバイス・結果がアルゴリズムによる意思決定の結果であるかどうかについて、エンドユーザや主体に混乱をもたらす可能性がありますか？
  - エンドユーザや主体は、意思決定やコンテンツ・アドバイス・結果がアルゴリズムによる意思決定の結果であることを適切に認識していますか？
- AIシステムは、人間とAIシステムのどちらと相互作用しているかどうかについて、エンドユーザや主体に混乱を生じさせる可能性がありますか？
  - エンドユーザまたは主体は、AIシステムと相互作用していることを通知されていますか？
- AIシステムは、エンドユーザのAIへの過度の依存を生み出すことにより、人間の自律性に影響を与える可能性がありますか？
  - エンドユーザがAIシステムに過度に依存することを避けるための手順を導入しましたか？

信頼できるAIのためのアセスメントリスト（項目例）

- AIシステムは、その他の意図しない、また望ましくない方法でエンドユーザの意思決定プロセスを妨害することにより、人間の自律性に影響を与える可能性がありますか？
  - AIシステムが誤って人間の自律に影響を与えないようにするための手順を実行しましたか？
- AIシステムは、エンドユーザや主体との、またはそれらの間の社会的相互作用をシミュレートしていますか？
- AIシステムは、人間の愛着を生み出したり、中毒性の行動を刺激したり、ユーザの行動を操作したりするリスクがありますか？起こりうるリスクまたは起こりそうなリスクに応じて、以下の質問に答えてください。
  - AIシステムへの不相応な愛着が生じた場合に、エンドユーザや主体に起こり得るマイナスの結果に対処するための対策を講じましたか？
  - 中毒のリスクを最小限に抑えるための対策を講じましたか？
  - ユーザ行動の操作のリスクを軽減するための対策を講じましたか？

信頼できるAIのためのアセスメントリスト（項目例）

○要件3．プライバシーとデータガバナンス

◇ プライバシー

- このサブセクションでは、プライバシーとデータ保護へのAIシステムの影響を自己評価する。プライバシーとデータ保護は、互いに密接に関連した基本的人権であり、個人の精神的・身体的な完全性をカバーする。

- AIシステムがプライバシーの権利、身体的・精神的・道徳的な完全性（不可侵性）の権利、データ保護の権利に与える影響を考慮しましたか？
- ユースケースに応じて、AIシステムにおけるプライバシー関連の問題を報告できるメカニズムを確立しましたか？

◇ データガバナンス

- このサブセクションでは、データ保護に関する様々な要件をAIシステムが遵守しているかを自己評価する。

- AIシステムは、個人データ（特別な種類の個人データを含む）を利用したり処理することによって、学習したり開発されていますか？
- 一般データ保護規則（GDPR）で義務付けられている以下の措置、または欧州以外の同等の措置を講じていますか？

信頼できるAIのためのアセスメントリスト（項目例）

- データ保護影響評価（DPIA）。
- データ保護責任者（DPO）を指名し、AIシステムの開発、調達、利用フェーズの早い段階で関与させているか。
- データ処理の監視メカニズム（アクセス制限、データアクセスと変更を記録するメカニズムを含む）。
- プライバシーバイデザインとバイデフォルトを達成するための措置（例：暗号化、仮名化、アグリゲーション、匿名化）。
- データの最小化、特に個人データ（特別な種類の個人データを含む）の最小化

  - AIシステムの開発に、同意を撤回する権利、異議を述べる権利、忘れられる権利を実装しましたか？
  - AIシステムのライフサイクルの過程で収集、生成、または処理されたデータのプライバシーとデータ保護への影響を考慮しましたか？

- AIシステムの学習などで用いられた非個人データやのプライバシーおよびデータ保護への影響を考慮しましたか？
- AIシステムを関連する標準（ISO、IEEEなど）や、データ管理やガバナンスで広く採用されているプロトコルに適合させましたか？

欧州委員会のAIホワイトペーパー

- 欧州委員会は2020年2月19日に、「AIに関するホワイトペーパー：優越と信頼に向けた欧州アプローチ」（White Paper on Artificial Intelligence - A European approach to excellence and trust） を公表。

  （https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en）

- 同ホワイトペーパーは、EUがAIに関する独自の「優越のためのエコシステム」と「信頼のためのエコシステム」を構築することを目標として掲げ、その実現に向けた政策オプションを提示。

- 前者の「優越のためのエコシステム」については、欧州産業および専門家市場の強化に向け、加盟国や民間部門と協力して、研究やイノベーションにおける投資を拡大し、AIスキル開発を強化し、中小企業等におけるAI導入にインセンティブを与えることを提案している。この目標のため、今後10年間にわたり、民間資金も合わせ、年間200億ユーロをAIに投資する方針。

- 後者の「信頼のためのエコシステム」については、AIが基本的人権や安全性に及ぼしうるリスクについて指摘した後、ハイリスクなAIアプリケーションについては、6つの義務的要件を含む新たな規制フレームワークが必要とし、更なる検討を行うとしている。また、ハイリスクなAIアプリケーションに対しては事前適合性評価が提案され、それ以外のAIアプリケーションに対しては自主的なラベリング制度が提案。

- 公共空間での自動顔認識（リモート生体識別）もハイリスクAIアプリケーションの1つとされているが、これについては特に1節を割いて言及がなされ、このようなAI利用が許可される条件について明確化するために、広範な関係者を巻き込んだ議論の場を立ち上げるとしている。

- 同ホワイトペーパーは、2020年6月14日（※）までパブリックコメントに付された。

  ※ 新型コロナの影響で2回にわたり締切が延長された。

欧州委員会のAIホワイトペーパー： リモート生体識別

IISE

〇ハイリスクなAIアプリケーションに課すべき要件

ｆ） リモート生体識別（自動顔認識など）に対する特別な要件

- リモートでの個人識別の目的での生体データの取得や利用は、例えば公共の場での顔認識のディプロイメントを通じたものは、基本的人権に特別なリスクをもたらす。リモート生体識別AIシステムを利用することの基本的人権への影響は、当該利用の目的、コンテキスト、範囲によってかなり変わりうる。
- EUのデータ保護ルールは、特定の条件下を除き、自然人をユニークに識別する目的での生体データの処理を原則として禁止している。とりわけ、GDPRの下では、そのような処理は限定された根拠に基づいてのみ行うことができる。そのうち主たる根拠は、重要な公共の利益である。そのようなケースでは、当該処理はEU法または加盟国の国内法に基づいて、比例性、データ保護の権利の尊重、および適切な安全管理措置の要件の下で、実施されなければならない。警察・刑事司法データ保護指令の下では、そのような処理に対する厳密な必要性や、原則としてEU法や国内法の規定、適切な安全管理措置がなければならない。自然人をユニークに識別する目的での生体データの処理はEU法で規定された禁止の例外に当たるため、EU基本権憲章の対象となる。
- EUの現行のデータ保護ルールやEU基本権憲章と整合的に、AIはリモート生体識別の目的では、そのような利用が十分に正当化され、比例的であり、十分な安全管理措置に服している場合にのみ、利用することができる。
- そのような公共の場でのAI利用に関連した社会的懸念に対処するために、またEU内部市場の断片化を避けるために、欧州委員会はそのようなAI利用（公共の場でのリモート生体識別の目的でのAI利用）を正当化するような特定の条件について、また共通の安全管理措置について、広範な欧州の議論を立ち上げる予定である。

AI白書のリモート生体識別（顔認識等）に対するEDPS見解

- EDPS（European Data Protection Supervisor：欧州データ保護監察官）は2020年6月29日に、欧州委員会のAIホワイトペーパーに対する意見「EDPS Opinion on the European Commission's White Paper on Artificial Intelligence」を公表。
（https://edps.europa.eu/sites/edp/files/publication/20-06-19_opinion_ai_white_paper_en.pdf）
- EDPSは同意見書で、AIホワイトペーパーで挙げられたリモート生体識別（Remote biometric identification：RBI）に関して、EUにおける公共空間での利用を一時的に禁止する考えを支持。
- 以下、同意見書からの抜粋。

- 62. 欧州委員会のAIホワイトペーパーは、リモート生体識別（RBI）によってもたらされる基本的権利に対するリスクを認識しており、これはEDPSも共有する見解である。リモート生体認証は2つの問題を生じさせる。1つは（遠くからの、大規模で、時には秘密での）個人の識別であり、もう1つは（遠くからの、大規模で、時には秘密での）生体データの処理である。AIに依存しているか否かに関わらず、これら2つの特徴のいずれかに関連する技術は同様な問題があり、RBIと同じ制限を適用する必要があるかもしれない。
- 63. RBIシステムによって個人の権利と自由にもたらされるリスク、たとえば公共の場でのリアルタイム顔認識のリスクは、適切に特定および軽減されなければならず、そのようなプロセスには、そのような技術の利用によって最も影響を受ける人々が含まれるべきである。RBIのリスクの一部は、RBIシステムが簡単に隠され、摩擦がなく、多くの場合は単なる「実験」として提示されるが、ユビキタスで広範にわたる監視システムに簡単に変えることができるという事実から生じる。

AI白書のリモート生体識別（顔認識等）に対するEDPS見解

- 64. RBIをサポートするインフラがひとたび整うと、簡単に他の目的で利用することができる（「ファンクションクリープ：機能の目的外利用」）。最近ではRBIシステムまたは他の技術インフラの一部を使用して、ソーシャルディスタンシングの測定やマスク使用の測定、温度チェック（カメラに温度計が統合されている場合）など様々な方法でパンデミックと戦うことができると主張する人もいる。これらの新しいアプリケーションの一部はGDPRの適用範囲に含まれない場合があるが、それでも民主主義社会では委縮効果をもたらすかもしれない。したがって、このようなAIの利用、およびそのようなファンクションクリープは、AIに関する規制で適切に対処するべきである。
- 65. RBIは重大な基本的権利の問題を生じさせる可能性があるが、EDPSは、個人の識別を目的としないRBI関連の技術も深刻なプライバシーの懸念を引き起こすことを強調したい。たとえば、リアルタイム顔認識に基づく感情検出では個人の感情を推測することができる。
- 66. このような技術がディプロイされる状況で、技術が必要であるか、比例的であるかどうか、またはそれが望まれているかどうかを評価することは最も重要である。EDPSはこの目的のために、情報に基づいた民主的な議論が行われるように、またEUや加盟国が特定のユースケースに対する各技術やシステムの比例性を保証するための包括的な法的枠組みを含めすべての適切な保護措置を講じるまで、EUにおける公共空間での自動認識のディプロイメントに対するモラトリアム（一時的禁止）の考えを支持する。それらには顔、歩容、指紋、DNA、声、キーストローク、その他のバイオメトリクスまたは行動信号などの人間の特徴の自動認識が含まれる。
- 67. 越境的・国内的な健康危機などの国家緊急事態の際の公的機関によるRBIの利用は、重要な公共の利益の理由で必要なものであり、EUまたは加盟国の法律に基づくものであり、透明性が高く、アカウンタブルであり、追求される目的に比例しており、特定の保護措置に服するものであり、期間が明確に制限されており、基本的人権の本質と人間の尊厳の尊重と矛盾しないもとであるべきである。

AI白書のリモート生体識別：　欧州委員会の動き

- 欧州委員会DG Connect（コミュニケーションネットワーク・コンテンツ・技術総局）のグロス課長（デジタル化産業課）は9月3日、欧州議会IMCO（域内市場・消費者保護委員会）の議員団に対し、「AI白書のパブリックコンサルテーションの結果を検討中であり、欧州委員会は欧州での顔認識技術の利用の将来的な禁止の選択肢を排除してはいない」と述べた。
- グロス課長は、GDPRは生体データの処理をカバーしうるが、欧州委員会は顔認識技術から取得したデータに関してGDPRの規定が十分であるかどうか、「追加の保護措置が必要かどうか、または特定ケース・特定エリアで一時的に顔認識を許可しない必要があるかどうかを検討する」と述べた。
- AI白書のパブリックコンサルテーションでは、公共空間でのリモート生体識別に対する懸念が多く寄せられ、グロス課長は欧州での利用について「EUでの特別な議論が必要」とした。同氏によれば、回答者の28%が公共の場でのリモート生体識別の禁止を支持しているという。
- 出典：Euractiv 2020年9月3日記事

（https://www.euractiv.com/section/data-protection/news/commission-will-not-exclude-potential-ban-on-facial-recognition-technology/）

米国顔認識ベンダーClearview AIに対するEDPB意見

- 米国のClearview AI は、FacebookやGoogle、Venmo、YouTubeなどのウェブサイトから（本人同意なく）取得された顔画像のデータベースを持ち、警察などの顧客から送信された容疑者などの顔画像をデータベースと照合し、一致した画像と出典元サイトへのリンク情報を提供するアプリケーションを提供している。データベースは30億枚以上の画像から構成され、米国FBIや英国国家犯罪対策庁など各国の600以上の法執行機関や、民間企業・学校・銀行などが利用しているという。
- 英国情報コミッショナーオフィス（ICO）とオーストラリア情報コミッショナー事務局（OAIC）は2020年7月に合同でClearview AIに対する調査を開始した。カナダのプライバシーコミッショナー事務所（OPC）も調査を開始し、7月に王立カナダ騎馬警察（RCMP）などカナダの全ての法執行機関が同社との契約を停止している。
- TwitterやGoogle、YouTube、Venmo、LinkedInなどの企業は相次いでClearview AIにデータの使用停止を求める通告書を送っている。民間企業による本人同意のない生体識別子の取得を禁じる生体情報プライバシー法（BIPA）のあるイリノイ州では、2020年1月に地方裁判所で同社が州民から訴訟を起こされている 。5月には米国の人権団体ACLU等からもイリノイ州のBIPAに違反したとして同州で訴訟を起こされている。ただし、同社は2020年8月に米国のICE（移民・関税執行局）と新たな契約を結んだという。
- （ニュースソース）
  - https://japan.cnet.com/article/35148187/
  - https://forbesjapan.com/articles/detail/35768
  - https://gigazine.net/news/20200707-clearview-ai-end-in-canada/
  - https://gigazine.net/news/20200207-clearview-ai-google-youtube-venmo-linkedin/
  - https://japan.cnet.com/article/35148516/
  - https://www.aclu.org/press-releases/aclu-sues-clearview-ai
  - https://jp.techcrunch.com/2020/08/15/2020-08-14-clearview-ai-ice-hsi-contract-2020/

米国顔認識ベンダーClearview AIに対するEDPB意見

- EDPB（欧州データ保護会議）は、欧州議員からのClearview AIに関する照会に対し、2020年6月10日に回答を行っている。
（https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_letter_out_2020-0052_facialrecognition.pdf）
- 同回答の中で、EU加盟国の政府機関による同社アプリケーションの利用は以下3つの観点から問題があるとしている。
- 1. EU法や加盟国法の下で構築されたデータベースを用いていない
- 「EDPBは、EU警察・刑事司法データ保護指令（EU）2016/680に基づいて、EU内の法執行機関が同指令第8条と第10条の厳格な条件に従ってのみ自然人を一意に識別する目的で生体データを処理できることに留意する。第8条によれば、そのような処理は、指令が適用される目的の下で、EU法または加盟国法に基づく業務の遂行に必要な範囲でのみ実行できる。また、それはEU基本権憲章や欧州人権条約を遵守するものでなければならない。第10条ではそのような処理は、とりわけ、厳密に必要であり、データ主体の権利と自由に対する適切な保護措置に服することが必要とされている。これらの厳格な条件に従って、EU内の法執行機関は、特定の状況下で、写真から得られた生体テンプレート（特徴データ）を含む生体データを処理し、公的機関の管理下にあり、EU法または加盟国法の下で構築されたデータベース内の生体テンプレートと照合を行うことが許される。
- しかし、法執行機関がClearview AIから提供されるようなサービスを利用することは、警察または犯罪捜査の一環として、個人データをEU域外の民間企業と共有することや、そのような民間企業の大規模で恣意的に構築されたオンラインアクセス可能な顔画像のデータベースに対して生体データの照合を行うことを意味するという点で、根本的に異なるものである。
- EDPBは、EU法または加盟国法がClearview AIから提供されるようなサービスを利用するための法的根拠を提供するかどうかについて疑問を持っている。したがって、現状のままで、将来の調査または保留中の調査を害することなく、EU内の法執行機関によるそのような利用の適法性を確証することはできない。」

米国顔認識ベンダーClearview AIに対するEDPB意見

- **2. 無差別かつ不正確なデータベースの利用は厳密な必要性と比例性の要件を満たさない**
- 「次にEDPBは、大量の個人データを何らの制限もなく、あるいはデータ間の正確な結合を行わずに、無差別に収集して構築されたデータベースに依存するような法執行コンテキストでの個人データの処理とその目的は、指令における厳密な必要性の要件を満たさない可能性が高いと考える。この比例性の原則の遵守に関して、EUレベルの私生活を尊重する基本権の保護は、欧州司法裁判所の判例に従い、厳密に必要な場合に限って個人データ保護の例外と制限を適用すべきであることを求めるものである。」

- **3. EU内の法執行機関から米国の民間企業への越境データ移転に該当する**
- 「最後にEDPBは、EU内の法執行機関が、Clearview AIなどのEU域外の拠点で活動を行うデータ管理者によって利用可能とされたアプリケーションを利用することは、EUから米国（同社の拠点がある国）への個人データの移転（例えば顔認識サービスを利用してアイデンティティが確認される人物の個人データの移転）を構成しうることに留意する。EDPBは特に、当該個人データの移転は、EU-USプライバシーシールド十分性決定の規定や、EU-USアンブレラ協定の対象とならないことに留意する。そのような移転が適法であるためには、EU内の法執行機関から第三国の民間企業への移転を具体的に規定する、EU警察・刑事司法データ保護指令第39条に定められた厳格な条件と要件を遵守しなければならないだろう。」

英国のAI倫理・データ保護関連の報告書・ガイダンス類

- ICO（情報コミッショナーオフィス）：
  - 「ビッグデータ、AI、機械学習とデータ保護（Big data, artificial intelligence, machine learning and data protection ver. 2.2）」（2017年9月改訂）
  - 「AIを用いた意思決定を説明する:ガイダンス（Explaining decisions made with AI: guidance）」（Alan Turing Instituteとの共同）（2020年5月20日）
  - 「AIとデータ保護に関するガイダンス（Guidance on AI and data protection）」（2020年7月30日）

- 英国庶民院科学技術委員会：
  - 「意思決定におけるアルゴリズム（Algorithms in decision-making）」（2018年5月）
- 英国貴族院AI特別委員会：
  - 「英国におけるAI：英国はAIを活用し、そして活用できる準備ができているか（AI in the UK: ready, willing and able?）」（2018年4月16日）
- データ倫理イノベーションセンター：
  - 「オンラインターゲティング」最終報告書（2020年2月）
  - 「アルゴリズムを用いた意思決定におけるバイアス」中間報告書（2019年7月）

ICOのAIとデータ保護に関するガイダンス

- 英国ICO（情報コミッショナーオフィス）は2020年7月30日に、「Guidance on AI and data protection（AIとデータ保護に関するガイダンス）」を公表。

  （https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-ai-and-data-protection/）

- 企業の実施するAIプロジェクト（開発やディプロイメント）に、データ保護の原則をどのように適用すればよいかを解説したガイダンス。
- ICOの「ビッグデータ、AI、機械学習とデータ保護」（2017年9月）と「AIを用いた意思決定を説明する：ガイダンス」（アランチューリング研究所と共同）（2020年5月）を補完するもの。
- 目次は以下。
  - 情報コミッショナー序文
  - エグゼクティブサマリー
  - ガイダンス概要
  - 第1部：　AIの説明責任とガバナンスへの影響は何か？
  - 第2部：　AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？
  - 第3部：　AIにおけるセキュリティとデータ最小化をどのように評価するべきか？
  - 第4部：　AIシステムにおいて個人の権利をどのように保証するか？
  - 用語集

ICOのAIとデータ保護に関するガイダンス： エグゼクティブサマリー

- 人工知能（AI）のアプリケーションは、私たちの生活の多くの側面にますます浸透している。我々はAIがもたらす明確なメリットだけでなく、AIが個人の権利と自由にもたらしうるリスクも理解している。
- そのためICOは、組織が自らAIシステムを設計する場合であれサードパーティから調達する場合であれ、データ保護遵守のベストプラクティスに焦点を当てて、「AI監査フレームワーク」を開発してきた。そのフレームワークは、AIアプリケーションを監査し、個人データを公平に処理するための明確な方法論を提供する。それは以下を含む：
  - ICOが監査および調査で使用する監査ツールおよび手順。
  - AIとデータ保護に関する本ガイダンス。
  - 自社AIシステムのコンプライアンスを監査する組織に実用的サポートを提供するためのツールキット（近日公開）。
- 本ガイダンスは、次の2つの関係者を対象とする。
  - データ保護責任者（DPO）、法務顧問、リスクマネージャー、上級管理職、ICOの監査人など、コンプライアンス関係者。
  - 機械学習の専門家、データサイエンティスト、ソフトウェア開発者とエンジニア、サイバーセキュリティとITリスクマネージャーなどの技術的専門家。
- 本ガイダンスでは、データ保護の観点からAIがもたらしうる権利と自由に対するリスクの評価方法と、それらのリスクを軽減するために実装できる適切な対策を明確にした。
- データ保護と「AI倫理」は重複する部分もあるが、本ガイダンスでは、AIを使用するための一般的な倫理原則や設計原則は提供していない。本ガイダンスはデータ保護の原則に対応し、次のような構成である。
  - 第1部： データ保護影響評価（DPIA）を含む、AIの説明責任（アカウンタビリティ）とガバナンス
  - 第2部： 適法性の根拠、AIシステムの性能評価と改善、潜在的な差別の軽減など、公平、適法で、透明な処理
  - 第3部： データの最小化とセキュリティ
  - 第4部： 自動意思決定に関連する権利を含む、個人の権利の遵守

米国NISTのAIワークショップシリーズ

- NIST（米国標準技術研究所）は、AIが米国の民間と公共部門におけるイノベーションを促進するものと捉え、AIの将来を実現するために必要な研究・標準・データに貢献。信頼できるAIシステムのイノベーションと導入を促進するには、AIシステムについての理解を深め、コミュニケーションし、信頼を高めることが必要とする。信頼できるAIシステムを実現するために、利害関係者は、信頼できるAIシステムのビルディングブロックを開発し、AI開発・使用・監視時にそれらのビルディングブロックを実装するための測定（measurement）・方法（method）・標準・ツールも開発する必要がある。NISTは、開発者とユーザのAIコミュニティと協力して、信頼できるAIに必要な測定・標準・関連ソリューションの開発に注力している。
（https://www.nist.gov/topics/artificial-intelligence）
- NISTはAIコミュニティと協力して、AIシステムが正確で確実性があり、安全でセキュアで、説明可能で、偏見がないという信頼（trust）を育むために必要な技術要件を特定中。コラボレーションを促進し、信頼できるAIの構成要素（ビルディングブロック）についての共通の理解を深めるために、NISTは政府・産業界・学界・その他の利害関係者をまとめた一連のワークショップを開催。ワークショップの焦点はAI標準と関連ツールの開発を進めることにある。

- 第1回ワークショップ「Exploring AI Trustworthiness」：2020年8月6日
  - 信頼できるAIシステムのビルディングブロックと、AI開発・使用・監視の際にそれらのビルディングブロックを実装するための測定・方法・標準・ツールについて議論のキックオフ。
- 第2回ワークショップ「Bias in AI」：2020年8月20日
  - AIのバイアスとは何か、そしてそれをどのように測定するかについての共有理解の促進。

米国NISTのAIワークショップシリーズ： 第2回のアジェンダ

9:00 AM	**Welcome and Introduction** – Elham Tabassi, Acting Chief of Staff, NIST Information Technology Laboratory
9:15 AM	**Plenary: Panel Session: Foundational Juggernaut: Addressing data bias challenges in AI** ▪ Darrell West *(Moderator)* Brookings Institution ▪ Andrew Burt, BNH, Immuta ▪ Alexandra Chouldechova, Partnership on AI, Carnegie Mellon University ▪ Fernando Diaz ▪ Teresa Tung, Accenture Labs
10:45 AM	**Breakout Session #1 – Data Bias** •What is the “right data”? •What are the biggest barriers to success at this point? •How can technology developers and practitioners effectively work together and inform each other to mitigate data bias? •From what shared understandings or assumptions about bias in AI do we need to operate before moving onto the next step? •What do we need in order to measure bias within data? •How does bias exhibit itself in data? Are there particular indicators?
11:50	**Plenary: Report Outs from Breakout Sessions - Data Bias**
12:30 PM	*1-hour Lunch Break*
1:30 PM	**Plenary: Panel Session: Algorithmic bias is in the question, not the answer: Measuring and managing bias beyond data** ▪Joshua Kroll *(Moderator)*, Naval Postgraduate School ▪ Aylin Caliskan, The George Washington University ▪ Abigail Jacobs, University of Michigan ▪ Nicol Turner Lee, Brookings Institution ▪ Kush Varshney, IBM
2:45 PM	**Breakout Session #2 – Bias in Algorithmic Modeling** •Algorithms are highly dependent on data, which can be biased. But how does the algorithmic process insert additional biases? •Is it possible to use algorithms to mitigate data biases? •What do we need in order to measure algorithmic bias? •How does bias exhibit itself in the model? •How can algorithms exacerbate biases? How can they be used to mitigate biases? •How can we think about transparency and its role in bias in AI?
3:50 PM	**Plenary: Report Outs from Breakout Sessions - Algorithmic Bias**
4:30 PM	Closeout (Elham Tabassi, NIST Information Technology Laboratory)

米国NISTの「説明可能なAIの4つの原則」ドラフト

- NISTは2020年8月に「説明可能なAIの4つの原則（NISTIR 8312）」のドラフトを公表。8月17日から10月15日までパブリックコメント。

  （https://www.nist.gov/topics/artificial-intelligence/ai-foundational-research-explainability）
- 説明可能なAIシステムの4原則を示すが、これらの原則はAIシステムと（説明を受ける）人間との相互作用に大きく影響される。すなわち、AIのアプリケーション・タスク・説明を受ける人間によって、適切とみなされる説明のタイプは異なる。
- 説明可能なAIの4つの原則
  - 説明（Explanation）：システムは全てのアウトプットに対して証拠または理由を提示すること。
  - 有意味性（Meaningful）：システムは個人ユーザが理解できる説明を提供すること。
  - 説明の正確性（Explanation Accuracy）：説明はアウトプットを生み出すシステムのプロセスを正確に反映していること。
  - 知識の限界（Knowledge Limits）：システムは設計された条件下でのみ、あるいはアウトプットが十分な信頼に達した場合のみ動作すること。
- 説明のタイプ
  - ユーザベネフィット（User benefit）
  - 社会的受容（Societal acceptance）
  - 規制とコンプライアンス（Regulatory and compliance）
  - システム開発（System development）
  - 所有者ベネフィット（Owner benefit）

米国NISTの「説明可能なAIの4つの原則」ドラフト

○ 説明（Explanation）の原則

- 「説明」の原則は、AIシステムに各アウトプットの証拠、サポート、または理由付けを提供するよう義務付けている。この原則自体は、証拠が正確で、インフォーマティブで、わかりやすいものであることを必要としない。この原則は、システムが説明を提供できることのみを示している。現在進行中の一連の作業が、説明可能なAIメソッドの開発と検証を目指している。現在、さまざまな戦略とツールがディプロイされ開発されている。この原則は、説明に対して品質の測定を課すものではない。有意味性の原則と、説明の正確性の原則が、説明を評価するためのフレームワークを提供する。

○ 有意味性（Meaningful）の原則

- 説明を受ける者がシステムの説明を理解すれば、システムは「有意味性」の原則を満たす。一般にこの原則は、ユーザが説明を理解できる場合、および/またはタスクを完了するのに役立つ場合に満たされる。この原則は、説明が1つのサイズですべてに当てはまることを意味するものではない。システムのユーザグループが複数ある場合は、異なる説明が必要になることがある。有意味性の原則により、各ユーザグループに合わせた説明が可能になる。ユーザグループは、システムの開発者vsシステムのエンドユーザ、弁護士/裁判官vs陪審員などとして広く定義できる。これらのグループの目標や切実な要求は異なる場合がある。たとえば、科学捜査官にとって有意味性は、陪審員にとって有意味性とは異なる場合がある。

米国NISTの「説明可能なAIの4つの原則」ドラフト

○ 説明の正確性（Explanation Accuracy）の原則

- 説明の原則と有意味性の原則だけでは、アウトプットを生成するシステムのプロセスを正しく反映した説明をシステムが提示することまでは求められない。「説明の正確性」の原則は、システムの説明にこのような正確性を課すものである。
- 説明の正確性は、意思決定の正確性とは異なる概念である。
- 様々なグループや個人ごとに異なる説明の正確性がありうる。専門家はシステムのプロセスを完全に記述するような詳細な説明を必要とするかもしれないが、より重要なポイントのみに焦点を当てた簡潔な説明を必要とするユーザもいるだろう。システムが複数のタイプの説明を生成できる場合、システムはより説明可能であると考えられる。

○ 知識の限界（Knowledge Limits）の原則

- 上記3つの原則は、システムがその知識の制限内で動作していることを暗黙的に前提している。この「知識の限界」の原則は、システムが動作するように設計または承認されていないケースや、アウトプットが信頼できないケースをシステムが識別していることを指す。知識の限界を識別して宣言することにより、不適切である可能性がある場合に判断が提供されないように、アウトプットを保護する。知識の限界の原則は、誤解を招く、危険な、または不当な意思決定やアウトプットを防止することにより、システムへの信頼を高めることができる。

米国NISTの「説明可能なAIの4つの原則」ドラフト

IISE

○5つの説明のタイプ

- ユーザベネフィット：この説明タイプは、アウトプットについてユーザに情報提供することを目的としている。たとえばこの説明は、ローン申請が承認された、または拒否された理由を提供する。
- 社会的受容：この説明タイプは、社会による信頼と受容を生み出すことを目的としている。たとえば予期しないアウトプットがシステムによって提供された場合、この説明は、なぜこのアウトプットが生成されたのかをユーザが理解するのに役立つ。また、根拠を提供できる場合は、システムの快適性を高めることができる。
- 規制とコンプライアンス：この説明タイプは、規制や安全基準などの遵守の監査に役立つ。説明の対象者には、重要な詳細を必要とするユーザ（例：安全性の規制者）やシステムとやり取りするユーザ（例： 開発者）が含まれるかもしれない。自動運転車の開発者や監査人が含まれるかもしれない。これには、飛行機墜落後の科学検査のアウトプットを評価するための説明も含れるかもしれない。
- システム開発：この説明タイプは、AIアルゴリズムやシステムの開発・改善・デバッグ・保守を支援・促進する。このカテゴリの説明の対象者には、技術スタッフ・製品マネージャー・経営幹部が含まれる。このカテゴリには、重要な詳細を必要とするユーザや、システムとやり取りするユーザが含まれる。たとえば、これには、勾配加重クラス活性化マッピング（GRAD-CAM）ベースのツールを使用して視覚アルゴリズムをデバッグする技術スタッフが含まれるかもしれない。
- 所有者ベネフィット：この説明タイプは、システムのオペレーターに利益をもたらす。 例えば、以前に閲覧したコンテンツに基づいて、次に視聴するべき映画やビデオをリストアップし、そのセレクションを説明するようなレコメンデーションシステムである。システムは映画を推奨し、「あなたはこれこれの映画が好きだったので、次に視聴するべき映画はこれです」と述べて、そのセレクションを説明する。ユーザは説明を信頼するとこのサービスで映画を視聴し続けるため、所有者に利益をもたらす。

（参考資料）

欧州委員会のAIホワイトペーパー概要

欧州委員会のAIホワイトペーパー

- 欧州委員会は2020年2月19日に、「AIに関するホワイトペーパー：優越と信頼に向けた欧州アプローチ」（White Paper on Artificial Intelligence - A European approach to excellence and trust） を公表。

  （https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en）

- 同ホワイトペーパーは、EUがAIに関する独自の「優越のためのエコシステム」と「信頼のためのエコシステム」を構築することを目標として掲げ、その実現に向けた政策オプションを提示。

- 前者の「優越のためのエコシステム」については、欧州産業および専門家市場の強化に向け、加盟国や民間部門と協力して、研究やイノベーションにおける投資を拡大し、AIスキル開発を強化し、中小企業等におけるAI導入にインセンティブを与えることを提案している。この目標のため、今後10年間にわたり、民間資金も合わせ、年間200億ユーロをAIに投資する方針。

- 後者の「信頼のためのエコシステム」については、AIが基本的人権や安全性に及ぼしうるリスクについて指摘した後、ハイリスクなAIアプリケーションについては、6つの義務的要件を含む新たな規制フレームワークが必要とし、更なる検討を行うとしている。また、ハイリスクなAIアプリケーションに対しては事前適合性評価が提案され、それ以外のAIアプリケーションに対しては自主的なラベリング制度が提案。

- 公共空間での自動顔認識（リモート生体識別）もハイリスクAIアプリケーションの1つとされているが、これについては特に1節を割いて言及がなされ、このようなAI利用が許可される条件について明確化するために、広範な関係者を巻き込んだ議論の場を立ち上げるとしている。

- 同ホワイトペーパーは、2020年6月14日（※）までパブリックコメントに付された。

  ※ 新型コロナの影響で2回にわたり締切が延長された。

はじめに

○はじめに

- AIはデータエコノミーの最も重要なアプリケーションの1つ。現在、ほとんどのデータは消費者に関するデータであり、クラウドベースのインフラに保存され処理されている。他方、今後いっそう増加するデータの大部分は産業やビジネス、公共部門から来るデータであり、様々なシステム、とりわけネットワークのエッジで機能するコンピューティング機器に保存されるだろう。このことは、デジタル化された産業やB2Bアプリケーションでは強いが、消費者プラットフォームでは比較的弱い欧州にとって新たな機会をもたらす。
- AIに対する欧州共通的なアプローチは、単一市場が十分な規模に到達し、単一市場の断片化を避けるために必要。各国単位でのイニシアティブの導入は、法的確実性を脅かし、市民の信頼を弱め、ダイナミックな欧州産業の出現を妨げるリスクがある。
- このホワイトペーパーは、欧州における信頼できるセキュアなAI開発を可能とするための政策オプションを提示するもの。ホワイトペーパーの主要な構成要素は以下の2つ。
    - バリューチェーン全体において「優越のためのエコシステム」を達成するために、欧州、国、地方レベルでの取り組みの整合を図る手段を規定した政策フレームワーク。
    - 欧州独自の「信頼のためのエコシステム」を作るための、欧州におけるAIの将来的な規制フレームワークの主要素。

AIのもたらすリスク

○AIのもたらすリスク

- AIのもたらしうる害悪は物質的なもの（個人の安全性や健康、人命の損失、財産への損害）と非物質的なもの（プライバシーの喪失、表現の自由の権利への制限、人間の尊厳、雇用へのアクセス等での差別）の両者であるかもしれず、広範なリスクに関係しうる。規制フレームワークは、様々な潜在的な害悪のリスク（とりわけ最も重大なもの）をいかに最小化するかに集中するべき。
- AIの利用に関連した主要なリスクは、①基本的人権（個人データとプライバシーの保護、被差別を含む）を保護するために規定されたルールの適用に関わるものと、②（製品）安全性と（製造物）責任に関連するものである。

①個人データとプライバシー保護、被差別を含む、基本的人権に対するリスク

- AIを用いた意思決定等を理解したり、異議申し立てを行うことの困難さ。
- AIが政府機関等によって大規模監視の目的で利用されたり、雇用主によって従業員の行動を監視する目的で利用されること。
- AIが匿名化されたデータの再特定に利用されること。
- オンライン仲介事業者（プラットフォーマー）によって、ユーザの情報を優先させたり、コンテンツの調整を行うために利用されること。
- 人間の意思決定と同様なバイアスがAIに存在しうること。
- 不透明さ（「ブラックボックス効果」）、複雑性、予測不可能性、部分的に自律的なふるまいなど、AI技術の特徴は、既存のEU法令のルールの遵守を証明することを困難にしたり、それらの有効な執行を妨げる恐れがある。

②安全性に対するリスクと、（製造物）責任制度を有効に機能させること

- AI技術は、製品やサービスに埋め込まれた際に、利用者に新たな安全性のリスクを提示する恐れがある。例えば物体認識技術における欠陥の結果、自動運転車が道路上の対象物を間違って特定し、傷害や物的損害を伴う事故を引き起こしうる。
- 製造物責任指令の下では製造者は欠陥製品に起因する損害に対して責任を負うが、自動運転車のようなAIベースのシステムでは、当該製品に欠陥があること、そして発生した損害との間に因果的なつながりがあることを証明することが難しい。

ハイリスクなAIアプリケーション

○ハイリスクなAIアプリケーション

- AIアプリケーションは、以下のクライテリアの両方に合致している場合にハイリスクとみなされるべき。
  - 第一に、AIアプリケーションが、その典型的な特徴に鑑みて、重大なリスクが発生することが予想されるような分野で導入されている場合。これらの分野は、新たな規制フレームワークにおいて明示的かつ網羅的にリスト化されるべき。例えば、医療、交通、エネルギー、一部の公共分野などである。このリストは定期的にレビューされ、必要な場合には修正されるべき。
  - 第二に、AIアプリケーションが、上記分野において、重大なリスクを生じさせる可能性の高いやり方で利用されている場合。例えば、医療は一般的に関連する分野であるのはもっともだが、病院における予約システムにおける欠陥は、通常は、法的介入を正当化するような重大性なリスクは呈さないだろう。ある利用についてのリスクレベルのアセスメントは、影響を受ける者への影響に基づいてなされうる。例えば、個人や企業の権利に法的効果または同様に重大な影響を生み出すAIアプリケーションの利用、傷害・死亡・重大な物質的・非物質的損害のリスクを呈示するAIアプリケーションの利用、個人や法人が合理的に回避することのできない影響を生み出すAIアプリケーションの利用などである。
- AIに関する新たな規制フレームワークに含まれる義務的な要件は、原則として、これら2つのクライテリアに照らしてハイリスクと特定されたAIアプリケーションにのみ適用されるだろう。
- ただし、重大なリスクにより、ある目的でのAIアプリケーションの利用がハイリスクとみなされるような例外的なケースもあるかもしれない。例えば以下のケースでは、分野にかかわらず、義務的な要件が適用されうる。
  - 個人の重要性の観点および雇用の平等性に取り組むEUの観点から、雇用プロセスでのAIアプリケーションの利用は常に「ハイリスク」とみなされるだろう。消費者の権利に影響するようなアプリケーションについても検討が可能であろう。
  - リモート生体識別の目的でのAIアプリケーションおよびその他の侵害的サーベイランス技術の利用は、常に「ハイリスク」とみなされるだろう。

ハイリスクなAIアプリケーションに課すべき要件

IISE

○ハイリスクなAIアプリケーションに課すべき要件

- AIハイレベル専門家グループのAI倫理ガイドライン等を考慮に入れると、ハイリスクAIアプリケーションに対する要件は、以下の主要な要素から構成されうるだろう。
  - a) 学習データ（training data）
  - b) データと記録の保存（data and record-keeping）
  - c) （利用者に）提供される情報（information to be provided）
  - d) 堅牢性と正確性（robustness and accuracy）
  - e) 人間の監督（human oversight）
  - f) リモート生体識別の目的で利用されるものなど、特定のAIアプリケーションに対する特別な要件

a) 学習データ（Training data）

- AIシステムの学習に用いられるデータセットに関して、以下の要件が想定されうる。
  - AIシステムを用いた製品やサービスの利用が安全であることの合理的な保証を提供することを目的とした要件。EUの安全性のルール（既存および補足的ルール）における標準のセットに合致するもの。例えば、十分に広範であり、危険な状況を回避するために必要な全ての関連するシナリオをカバーするデータセットでAIシステムが学習されていることを保証する要件。
  - AIシステムの利用が、禁止された差別を含む結果を生み出さないことを保証することを目的とした合理的な措置を取るための要件。これらの要件には、とりわけ、十分な代表性を有するデータセットを利用する義務が含まれうる。特に、性別や民族、その他の禁止された差別の基盤となる属性が、これらのデータセットに適切に反映されていることである。
  - プライバシーや個人データが、AIを用いた製品やサービスの利用において十分に保護されていることを保証することを目的とした要件。GDPRや警察・刑事司法データ保護指令がこれらについて規制している。

ハイリスクなAIアプリケーションに課すべき要件

○ハイリスクなAIアプリケーションに課すべき要件（続き）

b) 記録とデータの保存（Keeping of records and data）

- この要件は、AIシステムによる問題のあるアクションや意思決定をトレースバックし、検証することを可能にする。この目的のため、新たな規制フレームワークは、以下を保存すべきと規定しうる。
  - AIシステムの学習やテストに用いられるデータに関する正確な記録。主要な特徴に関する記述と、いかにデータセットが選択されたかを含む。
  - 一定の正当化されたケースでは、データセット自体。
  - AIシステムの構築・テスト・検証に用いられたプログラミングおよび学習の手法・プロセス・技術に関する文書。安全性や、禁止された差別につながりうるバイアスを回避することに関連するものを含む。

c) 情報の提供（Information provision）

- AIの責任ある利用を促進し、信頼を構築し、必要な救済措置を容易化するためには、ハイリスクなAIシステムの利用に関する十分な情報がプロアクティブな仕方で提供されることが重要である。したがって、以下の要件が検討されうる。
  - AIシステムの能力と制約に関する明確な情報の提供を保証すること。とりわけ、システムの目的、システムが意図された通りに機能できる条件、目的を達成する上で期待される正確性のレベル。これらの情報は、システムのデプロイヤーにとってとりわけ重要であるが、監督機関やシステムの影響を受ける者にとっても関連するかもしれない。
  - これとは別に、市民は（人間ではなく）AIシステムとやりとりをする際に明確に情報提供を受けるべきである。EUのデータ保護法令には既にこの類のルールが含まれているが、前述の目的を達成するためには、追加的な要件が求められるかもしれない。そうであるならば、不必要な負荷は避けるべきである。そのため、例えば、市民にとってAIシステムとやり取りしていることが直接的に明白であるような状況では、そのような情報は提供する必要がない。また、提供される情報は客観的で、簡潔で、容易に理解できるものであることが重要である。情報が提供される仕方は、個別のコンテキストに合わせて調節されるべき。

ハイリスクなAIアプリケーションに課すべき要件

○ハイリスクなAIアプリケーションに課すべき要件（続き）

d) 堅牢性と正確性（Robustness and accuracy）

- AIシステムは責任ある仕方で、発生するリスクについて事前に十分かつ適切な検討を行った上で、開発される必要がある。AIシステムの開発と機能は、それらが意図された通りに信頼できるふるまいを行うことを保証するものでなければならない。害悪のリスクを最小化するために、全ての合理的な措置が取られるべきである。したがって、以下の要素が検討されうる。
  - 全てのライフサイクルを通じて、AIシステムが堅牢で正確であること、あるいは少なくとも正確性のレベルを正しく反映していることを保証する要件。
  - 結果が再現可能であることを保証する要件。
  - 全てのライフサイクルを通じて、AIシステムが十分にエラーや不整合に対処できることを保証する要件。
  - データまたはアルゴリズム自体を操作しようとする明白な攻撃および隠れた試行に対して、AIシステムが回復力を有すること、またこれらのケースにおける軽減措置が取られていることを保証する要件。

e) 人間の監督（Human oversight）

- 人間の監督は、AIシステムが人間の自律を損ねたり、その他のマイナスの影響を及ぼさないことを保証する手助けになる。信頼できる、倫理的で、人間中心のAIの目標は、ハイリスクなAIアプリケーションでは人間の適切な関与を保証することによってのみ達成することができる。例えば、人間の監督は、網羅的ではないが、以下のような特徴を有しうる。
  - AIシステムのアウトプットは、人間によって事前にレビューされ確認されない限り、有効なものとはならない（例えば、社会保障給付申請の却下の判断は人間のみによって行われる）。
  - AIシステムのアウトプットは直ちに有効なものとなるが、その後の人間の関与は保障される（例えば、クレジットカード申請の却下の判断はAIシステムによって処理されるが、その後の人間のレビューが可能でなければならない）。
  - 運用中のAIシステムのモニタリングと、リアルタイムに介入したり動作を停止できること（例えば、自動運転車の運転が安全でないと人間が判断した時に、ストップボタンやストッププロセスを利用可能とする）。
  - 設計段階において、AIシステムに運用的な制約条件を課す（例えば、自動運転車がセンサーの信頼性が低下した場合など視界が悪い条件下での運転を停止する、あるいは先行車との一定の距離を維持する）。

リモート生体識別（自動顔認識など）に対する特別な要件

○ハイリスクなAIアプリケーションに課すべき要件（続き）

ｆ） リモート生体識別（自動顔認識など）に対する特別な要件

- リモートでの個人識別の目的での生体データの取得や利用は、例えば公共の場での顔認識のディプロイメントを通じたものは、基本的人権に特別なリスクをもたらす。リモート生体識別AIシステムを利用することの基本的人権への影響は、当該利用の目的、コンテキスト、範囲によってかなり変わりうる。
- EUのデータ保護ルールは、特定の条件下を除き、自然人をユニークに識別する目的での生体データの処理を原則として禁止している。とりわけ、GDPRの下では、そのような処理は限定された根拠に基づいてのみ行うことができる。そのうち主たる根拠は、重要な公共の利益である。そのようなケースでは、当該処理はEU法または加盟国の国内法に基づいて、比例性、データ保護の権利の尊重、および適切な安全管理措置の要件の下で、実施されなければならない。警察・刑事司法データ保護指令の下では、そのような処理に対する厳密な必要性や、原則としてEU法や国内法の規定、適切な安全管理措置がなければならない。自然人をユニークに識別する目的での生体データの処理はEU法で規定された禁止の例外に当たるため、EU基本権憲章の対象となる。
- EUの現行のデータ保護ルールやEU基本権憲章と整合的に、AIはリモート生体識別の目的では、そのような利用が十分に正当化され、比例的であり、十分な安全管理措置に服している場合にのみ、利用することができる。
- そのような公共の場でのAI利用に関連した社会的懸念に対処するために、またEU内部市場の断片化を避けるために、欧州委員会はそのようなAI利用（公共の場でのリモート生体識別の目的でのAI利用）を正当化するような特定の条件について、また共通の安全管理措置について、広範な欧州の議論を立ち上げる予定である。

法的要件の適用対象者

○法的要件の適用対象者

- ハイリスクなAIアプリケーションに対する法的要件の適用対象となる者については、以下の2つのイシューを検討するべきである。
- 第一に、関係する事業者にどのように義務が分担されるべきかの問題である。多くのアクターはAIシステムのライフサイクルに関与している。これには、開発者、デプロイヤー（AIを用いた製品やサービスを利用する者）、潜在的な関係者（生産者、ディストリビューターまたは輸入者、サービスプロバイダー、専門家、私的利用者）が含まれる。将来的な規制フレームワークにおいては、各々の義務は潜在的なリスクに対処するのに最もよい位置にあるアクターに割り当てられるべきである。例えば、AIの開発者は、開発段階で生じるリスクに対処するのに最もよい位置にあるかもしれないが、利用段階でのリスクをコントロールする能力は、開発者には限られているかもしれない。
- 第二に、新たな規制フレームワークによる法的介入の地理的な範囲に関する問題である。欧州委員会の見解では、AIを用いた製品やサービスをEU域内で提供する、全ての関連する経済的オペレーターは、EU域内に拠点を有するか否かに関わらず、規制フレームワークの要件が適用される（域外適用される）ということが肝要である。

事前適合性評価

○事前適合性評価

- AIアプリケーションが市民や社会に呈するハイリスクの観点からは、欧州委員会は現段階では、客観的な事前適合性評価によって、ハイリスクなAIアプリケーションに適用される前述の義務的な要件が遵守されていることを検証し保証することが必要だろうと考える。事前適合性評価は、テスト（testing）、検査（inspection）、あるいは認証の手続きを含みうる。また、事前適合性評価は、開発段階で用いられたアルゴリズムやデータセットのチェックを含みうる。
- 事前適合性評価に依存した制度を設計し実装する場合には、以下が考慮されるべきである。
  - 事前適合性評価では、前述の全ての要件を検証することが適切ではないかもしれない。例えば、情報提供の要件は一般的には、事前適合性評価を通じて検証するのに適していない。
  - ある種のAIシステムは経験によって進化し学習する可能性を考慮すべきである。この場合、AIシステムのライフサイクルに渡って、繰り返しの評価が必要になるかもしれない。
  - 学習やプログラミングに用いられるデータを検証することの必要性。またAIシステムの構築・テスト・検証に用いられるプログラミングおよび学習手法、プロセス、技術を検証することの必要性。
  - 事前適合性評価によってAIシステムが要件を満たしていないことが示された場合、例えば学習データに関して要件が守られていないことが示された場合、特定された欠陥は修正される必要がある。例えば、全ての要件に合致することを保証するような仕方で、EU域内の当該システムを再学習させる必要がある。
- 事前適合性評価は、その拠点の場所に関わらず、新たな規制フレームワークの要件を割り当てられたすべての事業者の義務となるだろう。中小企業に対する負荷を制限するためには、デジタルイノベーションハブを通じたものを含め、何らかのサポート構造が想定されるかもしれない。さらない、標準や専用オンラインツールもまた遵守を容易化しうるだろう。

自主的なラベリング制度

○自主的なラベリング制度

- 「ハイリスク」とみなされず、前述の義務的な要件に服しないAIアプリケーションについて、1つのオプションは、適用法令に加えて、自主的なラベリング制度を設立することである。
- このラベリング制度の下では、関心のある事業者は、前述の要件、またはラベリング制度のために特別に設定された同様な要件のセットに、ボランタリーベースで、自ら従うことを決めることができる。これにより、これらの事業者は、自らのAIアプリケーションに対する品質ラベルの資格を得ることができる。
- 自主的なラベルは、当該事業者オペレーターがAIを用いた自らの製品やサービスが信頼できることを表示することを可能とする。またこれらの製品やサービスが、通常の法的義務を超えて、一定の客観的で標準化されたEU全体のベンチマークを遵守していることを、利用者が容易に認識することを可能にする。このことは、AIシステムにおける利用者の信頼を高め、AI技術の全体的な普及を促進するに寄与する。
- このようなオプションは、ハイリスクとみなされないAIシステムの開発者および／またはデプロイヤーのための自主的なラベリングフレームワークを規定する新たな法的文書（legal instrument）の作成を伴う。このラベリング制度への参加は自発的であるが、開発者やデプロイヤーがラベル利用を選択した場合には、その要件は拘束的になるだろう。

（参考資料）

英国ICOのAIとデータ保護に関するガイダンス概要

ICOのAIとデータ保護に関するガイダンス

- 英国ICO（情報コミッショナーオフィス）は2020年7月30日に、「Guidance on AI and data protection（AIとデータ保護に関するガイダンス）」を公表。

  （https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-ai-and-data-protection/）

- 企業の実施するAIプロジェクト（開発やディプロイメント）に、データ保護の原則をどのように適用すればよいかを解説したガイダンス。
- ICOの「ビッグデータ、AI、機械学習とデータ保護」（2017年9月）と「AIを用いた意思決定を説明する：ガイダンス」（アランチューリング研究所と共同）（2020年5月）を補完するもの。
- 目次は以下。
  - 情報コミッショナー序文
  - エグゼクティブサマリー
  - ガイダンス概要
  - 第1部：　AIの説明責任とガバナンスへの影響は何か？
  - 第2部：　AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？
  - 第3部：　AIにおけるセキュリティとデータ最小化をどのように評価するべきか？
  - 第4部：　AIシステムにおいて個人の権利をどのように保証するか？
  - 用語集

ICOのAIとデータ保護に関するガイダンス： エグゼクティブサマリー

- 人工知能（AI）のアプリケーションは、私たちの生活の多くの側面にますます浸透している。我々はAIがもたらす明確なメリットだけでなく、AIが個人の権利と自由にもたらしうるリスクも理解している。
- そのためICOは、組織が自らAIシステムを設計する場合であれサードパーティから調達する場合であれ、データ保護遵守のベストプラクティスに焦点を当てて、「AI監査フレームワーク」を開発してきた。そのフレームワークは、AIアプリケーションを監査し、個人データを公平に処理するための明確な方法論を提供する。それは以下を含む：
  - ICOが監査および調査で使用する監査ツールおよび手順。
  - AIとデータ保護に関する本ガイダンス。
  - 自社AIシステムのコンプライアンスを監査する組織に実用的サポートを提供するためのツールキット（近日公開）。
- 本ガイダンスは、次の2つの関係者を対象とする。
  - データ保護責任者（DPO）、法務顧問、リスクマネージャー、上級管理職、ICOの監査人など、コンプライアンス関係者。
  - 機械学習の専門家、データサイエンティスト、ソフトウェア開発者とエンジニア、サイバーセキュリティとITリスクマネージャーなどの技術的専門家。
- 本ガイダンスでは、データ保護の観点からAIがもたらしうる権利と自由に対するリスクの評価方法と、それらのリスクを軽減するために実装できる適切な対策を明確にした。
- データ保護と「AI倫理」は重複する部分もあるが、本ガイダンスでは、AIを使用するための一般的な倫理原則や設計原則は提供していない。本ガイダンスはデータ保護の原則に対応し、次のような構成である。
  - 第1部： データ保護影響評価（DPIA）を含む、AIの説明責任（アカウンタビリティ）とガバナンス
  - 第2部： 適法性の根拠、AIシステムの性能評価と改善、潜在的な差別の軽減など、公平、適法で、透明な処理
  - 第3部： データの最小化とセキュリティ
  - 第4部： 自動意思決定に関連する権利を含む、個人の権利の遵守

ICOのAIとデータ保護に関するガイダンス： エグゼクティブサマリー

IISE

○第1部（AIの説明責任とガバナンスへの影響は何か？）について

- 説明責任の原則により、組織はデータ保護を遵守し、AIシステムにおける遵守を証明する責任がある。AIのコンテキストでは、説明責任を果たすために次のことが必要である。
  - AIシステムにおけるデータ保護の遵守に責任を負う。
  - リスクを評価して軽減する。
  - AIシステムがどのようにデータ保護を遵守しているかを文書化し、証明し、選択した措置を正当化する。
- これらのイシューは、使用する予定のAIシステムのDPIAの一部として検討する必要がある。個人データを処理するAIシステムを使用する場合、ほとんどの場合、DPIAの実施が法的に求められることに注意すべきである。DPIAは、AIシステムを使用して個人データを処理する方法と理由、および潜在的なリスクを検討する機会を提供する。
- また、管理者と処理者の関係を識別し、理解するように注意する必要がある。これは、AIサプライチェーンに通常含まれる様々の処理の複雑さと相互依存性によるものである。
- AIシステムのコンテキストで、データ保護の権利とその他の基本的人権の間に必要なバランスをとる一環として、必然的に、様々な競合する考慮事項と利益を考慮する必要がある。設計段階で、これらが何であるかを識別して評価する必要がある。次に、データ処理の目的と、それが個人の権利と自由にもたらすリスクのコンテキストで、それらをどのように管理できるかを決定する必要がある。ただし、AIシステムが個人データを処理する場合は、常に基本的なデータ保護の原則を遵守する必要があり、この要件を「トレード」することはできないことに注意するべきである。

○第2部（AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？）について

- AIを使用して個人データを処理する場合、AIが適法で、公平で、透明であることを確認する必要がある。これらの原則の遵守は、AIのコンテキストでは難しい場合がある。

ICOのAIとデータ保護に関するガイダンス： エグゼクティブサマリー

○第3部（AIにおけるセキュリティとデータ最小化をどのように評価するべきか？）について

- AIシステムは、既知のセキュリティリスクを増大させ、管理をより困難にする可能性がある。また、データ最小化の原則を遵守する上での課題も生じる。
- AIが増大させる可能性のある2つのセキュリティリスクは、次の2つである。
  - AIシステムの学習にしばしば必要となる大量の個人データの紛失や誤用。
  - 新しいAI関連のコードとインフラストラクチャの導入の結果としてもたらされるソフトウェアの脆弱性。
- デフォルトでは、AIを開発およびディプロイするための標準的な方法には大量のデータの処理が含まれる。これがデータ最小化の原則を遵守しないリスクがある。ただし、データの最小化と効果的なAIの開発・ディプロイメントの両方に役立つ多くの手法が存在する。

○第4部（AIシステムにおいて個人の権利をどのように保証するか？）について

- AIシステムの開発とディプロイの方法は、個人データが通常とは異なる方法で管理および処理されることを意味する。これにより、個人の権利がこれらのデータにいつどのように適用されるかを理解することが難しくなり、個人がそれらの権利を行使するための有効なメカニズムを実装することがより困難になる可能性がある。

第1部：AIの説明責任とガバナンスへの影響は何か？（概要）

○AIガバナンスとリスク管理にどのように取り組むべきか？

○意味のあるリスク選好をどのように設定するべきか？

○AIのデータ保護影響評価を実施する際に、何を考慮する必要があるか？

- データ保護法の下でDPIAが必要なのはなぜか？
- DPIAを行うか否かをどのように決定するか？
- DPIAで何を評価するべきか？
- データ処理をどのように記述するか？
- 誰かに相談する必要はあるか？
- 必要性と比例性をどのように評価するか？
- 個人に対するリスクをどのように特定し評価するか？
- リスク軽減策をどのように特定するか？
- DPIAをどのように結論付けるか？
- 次に何を行うか？

○AIにおける管理者／処理者の関係をどのように理解するべきか？

- AIシステムにとってなぜ管理者シップが重要か？
- 組織が管理者であるか処理者であるかをどのように判断するか？
- どのような種類の決定を行うと、組織は管理者とみなされるか？
- どのような種類の決定を行うと、組織は処理者とみなされるか？
- これらの問題を詳細に検討するためのICOの計画

第1部：AIの説明責任とガバナンスへの影響は何か？（概要）

○AI関連のリスクを評価する際に、競合する利益をどのように管理するべきか？

- AIの使用においては、データ保護法の要件を遵守する必要がある。ただし、考慮すべき様々な価値観や利益が存在する可能性があり、これらは時に競合する可能性がある。これらは「トレードオフ」と呼ばれ、適切なバランスは、その分野や社会的状況、また処理が個人に与えうる影響によって異なる。
  - 十分に正確なAIシステムをトレーニングすることの利益と、そのシステムをトレーニングするために処理される個人データの量を減らすことの利益
  - 十分に統計的に正確で、かつ差別を回避するAIシステムを作成すること
  - 説明可能性と、統計的正確性・セキュリティ・営業秘密の間の適切なバランスをとること

◆ これらのトレードオフをどのように管理できるか？

- AIシステムを設計または調達する際に、既存または潜在的なトレードオフを特定して評価し、それが個人に与えうる影響を評価する。
- トレードオフを最小限に抑えるために、利用可能な手法を検討する。
- 合理的なレベルの投資と労力で実装できる手法を検討する。
- 最終的なトレードオフの決定について、明確な基準と説明責任を持つ。
- 必要に応じて、個人や、AIアウトプットのレビュー担当者にトレードオフを説明するための措置を講じる。
- トレードオフを定期的に確認し、特に個人（またはその代表者）の見解と、それらを減らすための新しい手法やベストプラクティスを考慮に入れる。

◆ アウトソーシングおよびサードパーティAIシステム

◆ 文化、多様性、ステークホルダーとの関わり

◆ トレードオフを最小化するための数学的アプローチはどうか？

第2部：AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？（概要）

○適法性、公平性、透明性の原則はAIにどのように適用されるか？

- 第一に、AIシステムの開発とディプロイメントには、様々な目的のために様々な方法で個人データを処理することが含まれる。適法性の原則を遵守するためには、処理活動を個別の処理に分割し、それぞれについて目的と適切な適法性の根拠を特定しなければならない。
- 第二に、AIシステムを使用して個人データを推測する場合、この処理を公平に行うには、次のことを保証する必要がある。
  - AIシステムが十分に統計的に正確であり、差別を回避する。
  - 組織が個人の合理的な期待の影響を考慮する。
- 例えばローン返済率を予測するAIシステムは、次の場合に公平性の原則に違反する可能性がある。
  - しばしば不正確であることが判明するような予測を行う。
  - 正当な目的を達成するための比例的な手段として正当化することができないような、グループ間（例えば、男女間）の結果の不均衡につながる。
  - 個人が合理的に期待しない方法で個人データを使用する。
- 第三に、透明性の原則を遵守するために、AIシステムで個人データを処理する方法について透明性を保つ必要がある。AIと透明性の原則に関する主要な問題は、「AIを用いた意思決定を説明する」ガイダンスで扱われているため、本ガイダンスでは詳しく議論しない。

第2部：AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？（概要）

○AIを使用する場合、目的と適法性の根拠をどのように特定するか？

◆適法性の根拠を決定する際に何を考慮すべきか？

- 新しいAIシステムをトレーニングする場合でも、既存のAIシステムを使用して予測を行う場合でも、個人データを処理するときは常に、適切な適法性の根拠が必要である。
  - どの適法性の根拠がデータ処理に適用されるかを決定するのは組織の責任である。
  - 個人との関係の本質や処理目的を最もよく反映する適法性の根拠を常に選択しなければならない。
  - データ処理を開始する前に、この決定を行うべきである。
  - 決定を文書化するべきである。
  - 正当な理由がない限り、後日、適法性の根拠を変更することはできない。
  - プライバシー通知に（目的とともに）適法性の根拠を含めなければならない。
  - 特別なカテゴリのデータを処理する場合は、適法性の根拠と、（第9条の）追加条件の両方が必要である。

◆AIの開発目的とディプロイメント目的とをどのように区別するべきか？

- 多くの場合、処理目的と適法性の根拠を決定するときは、AIシステムの研究開発フェーズとディプロイメントフェーズを区別することが理にかなっている。したがって、AIの開発とディプロイメントに別個の適法性の根拠を選択する方が適切な場合がある。
  - AIシステムが汎用タスク用に開発された後、様々な目的で様々なコンテキストにディプロイされる場合、例えば顔認識システムのトレーニング後、その機能は、犯罪防止、認証、ソーシャルネットワークでの友達のタグ付けなどの複数の目的に利用されうる。これらの各アプリケーションは、異なる適法性の根拠を必要とするかもしれない。
  - サードパーティからAIシステムを調達する場合、サードパーティ開発者が行う個人データの処理は、システムの利用目的とは異なる目的（システムの開発など）で行われているため、別の適法性の根拠の特定が必要になる場合がある。
  - モデルのトレーニングを目的とした個人データ処理は個人に直接影響を与えないかもしれないが、モデルがディプロイされると、法的または重大な影響を与える自動意思決定を行う場合がある。その結果、開発段階とディプロメント段階で、利用可能な適法性の根拠が異なる場合がある。

第2部：AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？（概要）

◆ 同意に頼ることはできるか？

- AIに関連するデータ処理など、より複雑なデータ処理について有効な同意の取得を保証することは難しい場合がある。例えばデータの処理目的が多いほど、同意が本当に特定され、事前に説明を受けた上でのものであることを保証するのは難しくなる。
- 重要なのは、組織が個人データをどのように利用しているかを個人が理解し、この利用に同意しているということである。例えば様々なアウトプットを予測するために様々なモデルを探索して様々な属性を収集する場合、これらの処理活動について個人に通知し、有効な同意を取得するならば、同意が適切な適法性の根拠となるかもしれない。
- 同意はまた、AIシステムのディプロイメント中に個人データを利用するための適切な適法性の根拠となるかもしれない（例えばサービスのパーソナライズや、予測やレコメンデーションの作成などの目的）。
- ただし、同意が有効であるためには、個人が同意を与えたのと同じくらい容易に同意を撤回できなければならない。

◆ 契約の履行に頼ることはできるか？

- この適法性の根拠は、AIを使用した処理が、個人に契約的なサービスを提供するため、または契約締結前に個人の要求に応じて手段を講じるため（たとえば、AIによるサービス見積額を提供するため）に客観的に必要な場合に適用される。
- 同じサービスを提供できる、より侵害的でない方法がある場合や、当該処理が契約履行に客観的に見て必要でない場合、AIを使用したデータ処理についてこの適法性の根拠に頼ることはできない。

◆ 法的義務、公的職務、または生命に関する利益に頼ることはできるか？

- AIシステムを監査して、様々な法律を遵守していることを保証する必要がある場合もありうる。
- AIを公的な権限の行使の一部として使用する場合、または法律で定められた公共の利益のために職務を遂行する場合、個人データ処理はそれらの根拠に基づく場合がある。AIを使用して公共サービスを提供する公的機関など。
- AIシステムによる個人データ処理が、個人の生命に関する利益を保護することを根拠とする場合もありうる。例えば同意を提供できない患者の緊急医療診断である（AI診断システムによる意識不明患者のFMRIスキャンの処理等）。

第2部：AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？（概要）

◆ 正当な利益に頼ることはできるか？

- 正当な利益は最も柔軟な適法性の根拠である一方で、それが常に最も適切であるとは限らないことに注意することが重要である。たとえば、個人データを利用する方法が予期しないものであったり、不必要な害をもたらす場合である。それはまた、組織が個人の権利と利益を考慮し保護するための追加的責任を負っていることを意味する。また、データ処理の必要性と比例性を示すことができなければならない。
- 正当な利益には3つの要素があり、これらを「3部構成のテスト」と考えると分かりやすい。
  - 正当な利益を識別する（「目的テスト」）
  - それを達成するために当該処理が必要であることを示す（「必要性テスト」）
  - 個人の利益、権利、自由とバランスを取る（「バランステスト」）
- 様々な利益が「正当な利益」に該当しうる。これらは、組織自身や第三者の利益、また商業的利益や社会的利益でありうる。ただし重要なことは、正当な利益はより柔軟なものでありうる一方で、追加的責任が伴うことを理解することである。当該処理が個人に与える影響を評価し、当該処理に説得力のあるメリットがあることを示せる必要がある。
- 正当な利益の評価（LIA）の一環として、これらの考慮事項に対処し、文書化する必要がある。AIシステムの初期の研究開発段階では、目的は非常に広範な場合があるが、より具体的な目的が特定されたならば、それに応じてLIAをレビューする（または別の適法性の根拠を特定する）必要がある。

◆ 特別なカテゴリのデータと刑事犯罪に関するデータについてはどう扱うか？

◆ GDPR第22条の影響は何か？

- GDPR第22条には、個人に法的または同様に重大な影響を与える完全自動意思決定を実行している場合に、個人を保護するための追加的ルールが規定されている。AIシステムを使用してこのような意思決定を行う場合には、これが適用される。

第2部：AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？（概要）

○統計的正確性について何をする必要があるか？

◆データ保護法令における「正確性」とAIにおける「統計的正確性」の違いは何か？

- 「正確性」という言葉は、データ保護とAIでは異なる意味を持つことに注意することが重要である。データ保護における正確性は基本原則の1つであり、個人データが正確であり、必要に応じて最新の状態に保たれていることを保証する必要がある。処理する個人データが「事実に関して不正確または誤解を招く」ものではなく、必要に応じて、過度の遅滞なく修正または削除されるように、すべての合理的な措置を講じる必要がある。
- 他方、AI（より一般的には統計的モデリング）における正確性とは、AIシステムが正しくラベル付けされたテストデータに対して正しい回答を推測する頻度を指す。テストデータは通常、トレーニング前に学習データから分離されるか、別のソースから取得される。そして、しばしばAIシステムが提供する回答は個人データになる。
- 明確化のために、本ガイダンスでは次の用語を使用する。
  - 「正確性」は、データ保護法令における正確性の原則を指す。
  - 「統計的正確性」は、AIシステム自体の正確性を指す。
- データ保護における正確性の原則は、AIシステムへの入力として利用される個人データであろうと、AIシステムのアウトプットであろうと、すべての個人データに適用される。ただし、これは、AIシステムが正確性の原則を遵守するために100%統計的に正確である必要があることを意味するものではない。
- 多くの場合、AIシステムのアウトプットは、個人に関する事実情報として扱われることを意図されていない。代わりに、それらは、現在または将来の個人について真実でありうる何かについて、統計的に情報に基づいた推測を表すことを意図されている。このような個人データが事実と誤解されないようにするには、記録において事実ではなく統計的に情報に基づいた推測であることが示されていることを保証するべきである。この記録には、データの出所と、推測に用いられたAIシステムに関する情報も含めるべきである。
- AIシステムを使用して人々について推論する場合は、システムが目的に対して十分に統計的に正確であることを保証する必要がある。これは、すべての推論が正しい必要があるという意味ではないが、それらが正しくない可能性と、それらに基づいて行う意思決定に与える影響を考慮する必要がある。

第2部：AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？（概要）

◆ 様々な統計的正確性の指標をどのように定義し、優先順位を付ける必要があるか？

- 例えばAIシステムを使用して電子メールをスパム／非スパムとして分類する場合、統計的正確性の簡単な指標は、分析されたすべての電子メールに対する、スパムまたは非スパムとして正しく分類された電子メールの割合である。ただし、そのような指標は誤解を招く可能性もある。例として受信トレイに受信された電子メール全体の90%がスパムである場合、すべてをスパムとしてラベル付けすれば90%正確という分類を達成できてしまう。
- このため、代替的な統計的正確性の指標を使用して、システムがどれだけ優れているかを評価する必要がある。このような指標は、以下2つのタイプのエラー間のバランスを反映するべきである。
  - 偽陽性またはエラー「タイプI」： AIシステムが誤って陽性とラベルしてしまうケース（例えば非スパムの電子メールをスパムとして分類してしまう）。
  - 偽陰性またはエラー「タイプII」： 実際には陽性であるのに、AIシステムが誤って陰性とラベルしてしまうケース（例えば実際にはスパムである電子メールを非スパムとして分類してしまう）。
- これら2つのタイプのエラーの間で適切なバランスを取ることが重要である。次のような、これら2つのタイプのエラーを反映する有用な指標がある。
  - 適合率（precision）： 陽性と識別されたケースのうち実際に陽性であるケースの割合（「陽性予測値」とも呼ばれる）。例えばスパムとして分類された10通の電子メールのうち9通が実際にスパムである場合、AIシステムの適合率は90%である。
  - 再現率（recall）（または感度（sensitivity））： 実際に陽性であるケースのうち陽性と識別されたケースの割合。例えば100通の電子メールのうち10通が実際にスパムであるが、AIシステムがそのうち7通しか識別しない場合、その再現率は70%である。
- 適合率と再現率の間にはトレードオフがある。可能な限り多くの陽性ケースを見つけること（再現率を最大化すること）をより重要視する場合、これはいくらかの偽陽性のコスト（適合率の低下）をもたらす可能性がある。
- 例えば、人事採用の履歴書フィルタリングシステムが偽陽性を生み出す場合、資格のない志望者が面接に招待され、雇用主と志望者の時間を不必要に浪費する可能性がある。逆に偽陰性を生み出す場合は、資格のある志望者が雇用機会を逃し、雇用主は良い志望者を逃すだろう。リスクの重大度と性質に基づいて、どちらのエラーの回避を優先するか選択するべき。

第2部：AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？（概要）

○偏見や差別のリスクにどのように対処すべきか？

◆ AIシステムが差別につながりうるのはなぜか？

- 例：　或る銀行が見込み顧客の信用リスクを計算するためのAIシステムを開発している。銀行はAIシステムを使用して、ローン申請を承認または拒否する。このシステムは、職業、収入、年齢、ローンを返済したか否かなど、これまでの借り手に関するさまざまな情報を含む大規模なデータセットでトレーニングされている。銀行がテスト中に性別によるバイアスの可能性をチェックしたところ、女性の信用スコアが低くなる傾向があることを見出した。
- この場合、AIシステムは特定のグループ（女性）のメンバーを不利な立場に置くため、差別的であるように見える。このAIシステムが女性に低い信用スコアを与えるのには、様々な理由がありうる。
- 1つの理由は、偏った学習データである。学習データ内の性別の比率が偏っているかもしれない。たとえば、過去にローンを申請した女性が少なく、銀行に女性に関する十分なデータがないため、学習データに男性の借り手の割合が高くなっているかもしれない。
- AIシステムの作成に使用される機械学習アルゴリズムは、トレーニングやテストされるデータに最適になるように設計されている。男性が学習データで過大代表されている場合、モデルは男性の返済率を予測する統計的関係により多く注意を払い、女性の返済率を予測する統計的パターンにはあまり注意を払わない。言い換えると、統計的に「重要性が低い」ため、学習データセット内の女性が平均して男性よりもローン返済率が高い場合でも、モデルは体系的に女性のローン返済率を低く予測してしまうかもしれない。
- このような問題は、学習データで過小代表されているすべての集団に生じうる。たとえば、顔認識モデルが特定の民族と性別（例えば白人男性）に属する顔に偏ってトレーニングされている場合、そのグループの個人を認識するときはパフォーマンスが向上し、他のグループではパフォーマンスが低下するだろう。
- もう1つの理由は、学習データが過去の差別を反映している可能性があることである。例えば過去に、偏見のために女性からのローン申請が男性からのローン申請よりも頻繁に拒否された場合、そのような学習データに基づくモデルは、同じパターンの差別を再現する恐れがある。若い黒人男性に対する警察の職務質問や、伝統的に男性の役目とされる仕事の採用など、歴史的に差別が重大な問題となってきた領域では、この問題がより深刻に発生する可能性が高くなる。

第2部：AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？（概要）

- これらの問題は、学習データに性別や人種などの「保護すべき属性」が含まれていない場合でも発生する可能性がある。学習データの様々な変数（属性）は多くの場合、職業などの「保護すべき属性」と密接に関連している。これらの「プロキシ変数」により、設計者が意図していなくても、モデルはこれらの属性に関連する差別のパターンを再現することが可能である。
- これらの問題はどの統計モデルでも発生する可能性があるため、統計モデルを「AI」と見なさなくても、以下の考慮事項が当てはまるかもしれない。ただしAIシステムでは、より多くの属性（変数）を含めることができ、「保護すべき属性」のプロキシ変数となる属性の複雑な組み合わせを識別する可能性があるため、発生する可能性が高くなる。最新の機械学習手法の多くは、高次元データの非線形パターンを明らかにするのに優れているため、従来の統計的アプローチよりも強力であり、これらには差別を反映したパターンも含まれてしまう恐れがある。
- 他にも、AIシステムが差別的になる原因として以下が挙げられる。
  - 変数の測定・ラベル付け・集計方法における偏見またはバイアス
  - 開発者の偏った文化的背景
  - 不適切に定義された目的（例えば仕事の「最良の候補者」が性別、人種、その他の属性に関する前提を組み込んでいる場合）
  - モデルのディプロイ方法（たとえば、アクセシビリティ要件を満たしていないユーザーインターフェイスを介する場合）

◆ 機械学習モデルにおける差別リスクを軽減するための技術的アプローチは何か？

- 差別は技術だけでは現実的に「修正」することができない広範な問題だが、AI主導の差別を緩和することを目的としたさまざまなアプローチが存在する。コンピューター科学者等は、機械学習モデルが様々なグループの個人をどのように潜在的に差別的な方法で扱うか、またそれらを軽減するかを測定するために、さまざまな数学的手法を開発してきた。この領域は、アルゴリズム的「公平性」とよく呼ばれる。
- 「偏った学習データ」の例では、過小代表/過大代表された集団のサブセットについてデータを追加または削除することで（例えば女性からのローン申請に関するデータ数を追加することで）、バランスを取ることが可能でであるかもしれない。
- 「学習データが過去の差別を反映している」例では、データを修正するか、学習プロセスを変更するか、学習済みモデルを修正することで対処できる。

第2部：AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？（概要）

- これらの手法が効果的かどうかを測定するために、結果を測定できるさまざまな数学的「公平性」測定法がある。
- モデルが予測を行うために利用するインプットから「保護すべき属性」を単に削除するだけでは十分ではない可能性がある。「保護すべき属性」のプロキシ変数がしばしば存在するためである。他の測定法には、AIシステムが保護すべきグループ間でポジティブな結果やネガティブな結果（エラー）をどのように分配するかを比較することが含まれる。これらの測定法のいくつかは互いに矛盾している。すなわち、すべてを同時に満たすことはできない。これらの測定法のどれが、どのような組み合わせで最も適切であるかは、どのような組み合わせで、組織のコンテキストや適用法（例えば平等法）に依存する。
- また、AIシステムのパフォーマンスの統計的正確性に対するこれらの手法の影響も考慮する必要がある。例えば、差別の可能性を減らすために、信用リスクモデルを変更して、「保護すべき属性」が異なる人々（男性と女性など）間のポジティブな予測の割合が等しくなるようにすることができる。これにより、差別的な結果を防ぐことができるが、全体として統計的エラーの数が増える可能性もあり、これも管理する必要がある。
- ただし実際には、統計的正確性と差別の回避との間に常に緊張関係があるとは限らない。例えばモデルの差別的なアウトプットが、母集団内で統計的に少数のマイノリティに関するデータが相対的に不足していることによって引き起こされた場合には、マイノリティに関するより多くのデータを収集することによってモデルの統計的正確性を向上させ、なおかつ正しい予測の比率を平等にすることができるかもしれない。
- ただし、その場合、以下のような選択に直面するだろう。
  - マイノリティが直面する不均衡な統計的エラーの数を減らすために、マイノリティに関するより多くのデータを収集する。
  - マイノリティのデータを収集することのリスクのためにそのようなデータを収集しないことは、それらの個人の他の権利と自由にリスクをもたらす可能性がある。
- 状況によっては、統計的正確性や差別に関する他のセクター固有の規制義務もあるかもしれず、データ保護義務と一緒に検討する必要があるかもしれない。これらの義務を果たすために特定の方法でデータを処理する必要がある場合、データ保護はそれを妨げるものではない。

第2部：AIシステムの適法性、公平性、透明性を保証するために何をする必要があるか？（概要）

◆ AIシステムにおける差別を評価して対処するために、特別なカテゴリのデータを処理してよいか？

- AIシステムでの差別の可能性を評価して対処するには、2010年平等法で規定された「保護すべき属性」（年齢・障害・性転換・婚姻・同性婚・妊娠・出産・人種・宗教・信仰・性別・性的志向）を含むデータセットが必要になる場合がある。そのようなデータセットを用いて、保護すべき各グループでシステムがどのように機能するかをテストしたり、差別的な影響を回避するためにモデルを再トレーニングする可能性もある。
- このような分析を行う前に、そのような目的でデータを処理するための適切な適法性の根拠があることを確認する必要がある。

◆ 特別なカテゴリのデータ、差別、自動意思決定についてはどうすればよいか？

◆ AIを使用して、誤って特別なカテゴリのデータを推測してしまった場合はどうなるか？

- 居住地の郵便番号などの属性が、人種などの「保護すべき属性」のプロキシであるような多くのケースがある。ディープラーニングなどの機械学習の最近の進歩により、AIシステムは一見無関係なデータに反映されている「差別的なパターン」も簡単に検出できるようになってしまった。
- 例えば、採用決定で人間の意思決定者を支援するために志望者をスコアリングするために使用されるAIシステムは、以前に採用した従業員のサンプルでトレーニングされる場合がある。応募書類自体に含まれる情報には、人種、障害、メンタルヘルスなどの「保護すべき属性」が含まれていないかもしれない。ただし、トレーニングに使用された従業員のサンプルがこれらの理由で差別されていた場合（たとえば、業績評価で体系的に過小評価されていた場合）、アルゴリズムは、応募書類に含まれるプロキシデータからそれらの属性を推測することにより、設計者が意図せずとも、その差別を再現することを学習するかもしれない。
- AIシステムを使用する場合、モデルが「保護すべき属性」や特別なカテゴリのデータを推測している可能性を積極的に評価し、システムのライフサイクル全体でこの可能性を積極的に監視するべきである。システムが実際に特別なカテゴリのデータや有罪判決データを推測している場合（意図的かどうかにかかわらず）、第9条または第10条の条件を満たさなければならない。

◆ これらのリスクを軽減するためにどうすればよいか？

第3部：AIにおけるセキュリティとデータ最小化をどのように評価するべきか？（概要）

○AIはどのようなセキュリティリスクをもたらすか？

- セキュリティ要件は何か？
- 「従来の」テクノロジーと比較してAIにおけるセキュリティは何が異なるか？
- 学習データのセキュリティをどのように保証するべきか？
- この状況で私たちは何をすべきか？

○AIモデルにはどのような種類のプライバシー攻撃が適用されるか？

◆ モデル反転攻撃とは何か？

- モデル反転攻撃では、攻撃者が学習データに含まれる特定個人に関するデータの一部をすでに入手している場合、機械学習モデルの入力とアウトプットを観察することで、同一個人に関する追加的な個人データを推測できる。攻撃者が知ることができる情報は、同様の属性を持つ個人に関する一般的な推測を超えたものである。
- 例1–モデル反転攻撃：　初期の事例は、遺伝子バイオマーカーを含む患者データを使用して、抗凝固剤の正しい投与量を予測するように設計された医療モデルに関連するものである。学習データに含まれる個人に関する人口統計情報を入手した攻撃者が、基礎となる学習データ自体にはアクセスできないにもかかわらず、モデルから当該個人の遺伝子バイオマーカー情報を推測できることが証明された。
- 例2–モデル反転攻撃：　最近の別の事例は、攻撃者が顔認識技術（FRT）システムのトレーニングに使用された顔画像を再構成できることを示した。FRTシステムは、多くの場合、サードパーティがモデルを照会できるように設計されている。モデルに登録された人物の顔画像が送られると、モデルはその人物の氏名の推測結果と信頼率を返答する。攻撃者は、ランダムに生成されたさまざまな顔画像を送信することで、モデルを調査し、モデルから返された氏名と信頼率を観察することで、学習データに含まれる顔画像を再構成することができた。

◆ メンバーシップ推論攻撃とは何か？

- メンバーシップ推論攻撃により、攻撃者は、特定の個人が機械学習モデルの学習データ内に存在するか否かを推測できる。ただしモデル反転攻撃とは異なり、必ずしも当該個人の個人データ内容を推測するわけではない。

第3部：AIにおけるセキュリティとデータ最小化をどのように評価するべきか？（概要）

○AIモデルに対するプライバシー攻撃リスクを管理するためにどのような措置を取るべきか？

- 機械学習モデルがプライバシー攻撃に対して脆弱である1つの原因は、「過学習（overfitting）」として知られている。これは、モデルが学習データの詳細に過度に注意を払い、一般的なパターンだけでなく、学習データの個々のサンプルをほとんど覚えてしまっている場合である。過学習は、学習データにあまりに多くの属性が含まれている場合や、またはあまりに少ないサンプルしか用いられていない場合に発生する可能性がある。モデル反転攻撃やメンバーシップ推論攻撃は、これを悪用する可能性がある。
- 過学習を回避することは、プライバシー攻撃のリスクを軽減するのに役立つだけでなく、モデルがこれまでに見たことのない新たなサンプルについて適切な推論を行えるようにするのにも役立つ。ただし、過学習を回避しても、リスクが完全に排除されるわけではない。
- 上記のFRTの例のように、機械学習システムによって提供される信頼率を悪用できる場合は、それをエンドユーザーに提供しないことでリスクを軽減できるかもしれない。これは、通常のエンドユーザーが信頼率を知るニーズとバランスを取る必要があり、特定のユースケースに依存するだろう。

◆説明可能なAIによって引き起こされるAIセキュリティリスクについてどうすべきか？

- 最近の研究では、機械学習モデルを説明可能にするために提案されたいくつかの方法によって、モデルに対するプライバシー攻撃の実行が容易になってしまうことが示されている。例えば個人に説明を提供する場合、AIモデルがどのように機能するかについての独自情報が明らかになるリスクがある。ただし、商業的利益をデータ保護要件（たとえば、商業的セキュリティやデータ保護セキュリティ）と混同してはならない。
- AIについてデータ主体に提供する必要のある種類の説明は、「明確で平易な言葉を使用して、簡潔で、透明で、わかりやすく、簡単にアクセスできる形式である」必要があることを考えると、通常、商業的にセンシティブな（セキュリティ）情報を危険にさらすことはないだろう。ただし、個人の説明を受ける権利と（例えば）営業秘密を守る企業の利益との考慮が必要なケースもあるかもしれない（その場合でも、データ保護コンプライアンスは「トレードオフ」できない）。

◆敵対的サンプルに対してどう対処すべきか？

第3部：AIにおけるセキュリティとデータ最小化をどのように評価するべきか？（概要）

○AIシステムで利用できるデータの最小化とプライバシー保護技術は何か？

◆データ最小化原則について、どのような考慮が必要か？

- データ最小化の原則では、目的を達成するために必要な最小限の個人データを特定し、その情報のみを処理し、それ以上は処理しないようにする必要がある。ただし、AIシステムは一般的に大量のデータを必要とする。したがって、一見、AIシステムがデータ最小化の原則をどのように遵守できるかを理解するのは難しいかもしれないが、遵守する必要がある。

◆教師あり機械学習モデルで個人データをどのように処理するべきか？

- 「学習データ」を用いて教師あり機械学習アルゴリズムをトレーニングして、パターンを識別できるようにし、モデルを作成することができる。具体的には、学習データには、「ターゲット変数」（モデルが予測や分類することを目的としている変数）といくつかの「予測変数」（予測を行うために使用される入力）の両方が含まれる。
- 例えば銀行の信用リスク機械学習モデルの学習データには、予測変数には過去の顧客の年齢、収入、職業、場所が含まれ、ターゲット変数には顧客がローンを返済したかどうかが含まれる。
- トレーニングが完了すると、機械学習システムは、システムがこれまでに見たことのないサンプルを含む新しいデータに基づいて分類し、予測を行うことができる。クエリがモデルに送信され、新たなサンプルの予測変数（たとえば、新しい顧客の年齢、収入、職業）が含まれる。モデルは、この新たなサンプルのターゲット変数に関する最善の推測で応答する（例えば新しい顧客がローンを債務不履行にするどうか）。
- すなわち、教師あり機械学習アプローチでは、次の2つの主要なフェーズでデータを利用する。
  - １．学習フェーズでは、学習データは、過去のサンプルに基づいたモデルを開発するために利用される。
  - ２．推測フェーズでは、モデルを使用して新たなサンプルに関する予測や分類が行われる。
- モデルを使用して個々の人々に関する予測や分類を行う場合、学習フェーズと推論フェーズの両方で個人データが利用される可能性が非常に高くなる。

第3部：AIにおけるセキュリティとデータ最小化をどのように評価するべきか？（概要）

◆ 機械学習アプリケーション設計の際、個人データ最小化のためにどのような手法を利用するべきか？

◆ 学習段階で個人データを最小化するにはどうするべきか？

- 学習データに含まれるすべての属性が必ずしも処理目的に関連しているとは限らない。例えば、すべての金融属性や人口統計属性が信用リスクの予測に役立つとは限らない。したがって、どの属性（どのデータ）が目的に関連しているかを評価し、そのデータのみを処理する必要がある。
- モデルに含めるのに役立つ属性を選択するためにデータサイエンティストが使用するさまざまな標準的な属性選択方法がある。これらの方法は、データサイエンスではグッドプラクティスだが、データ最小化の原則を満たすためにも役立つ。

◆ データ最小化と統計的正確性のバランスをどのように取るべきか？

- 一般に、AIシステムが（機械学習のケースのように）データから学習する場合、学習データが多いほど、統計的に正確になる。つまり、データセット内の属性間に横たわる統計的に有用な関係を把握する可能性が高くなる。
- 例えば顧客の購入履歴に基づいて将来の購入行動を予測するモデルは、学習データに含まれる顧客が多いほど、統計的に正確になる傾向があるかもしれない。また、既存のデータセットに追加された新たな属性が、モデルの予測するデータに関連しているかもしれない。例えば購入履歴に加えて何らかの人口統計データを追加することで、モデルの統計的正確性が向上するかもしれない。
- ただし一般的に、各個人について収集されるデータ項目が多く、データセットに含まれるサンプル数が多いほど、特定の目的でデータが収集されたとしても、個人に対するリスクは大きくなる。データ最小化の原則では、目的に必要な以上のデータを利用しないようにする必要がある。したがって、含まれるデータ項目やサンプルの数が少なくても十分な正確性を達成できる場合は、そうするべきである。

◆ どのようなプライバシー強化手法を検討するべきか？

◆ 推測段階で個人データを最小化するにはどうするべきか？

◆ 匿名化には役割があるか？

◆ 学習データの保存と制限についてはどうするべきか？

第4部： AIシステムにおいて個人の権利をどのように保証するか？（概要）

○個人の権利はAIライフサイクルの様々な段階にどのように適用されるか？

◆ 学習データに対する個人の権利の要求をどのように保証するべきか？

- 学習データにおいて個人を識別することは、彼らの権利を保証する上での課題である。通常、学習データには、過去のトランザクション、人口統計情報、ロケーションなどの情報のみが含まれ、氏名・連絡先情報やユニークな顧客IDは含まれない。また学習データは、通常、機械学習アルゴリズムに適したものにするために、様々な加工がなされる。例えば顧客の購入の詳細なタイムラインは、トランザクション履歴のピークとトラフの要約に変換されるかもしれない。
- このようなプロセスでは、個人データを元の形式から詳細度の低い形式に変換するため、学習データを特定個人にリンクすることが非常に困難になる可能性がある。ただしデータ保護法では、これは必ずしもそのデータを範囲外にするのに十分であるとは見なされなず、個人の権利行使の要求に対応する場合は、このようなデータも考慮する必要がある。
- 当該データに識別子や連絡先詳細が含まれず、さらに前処理によって変換された場合でも、学習データは個人データと見なされる場合がある。これは、氏名が分からなくても、学習データにおいて個人を単独で（または他のデータと組み合わせて）「シングルアウト」できる場合である。例えば購入予測モデルの学習データに、1人の顧客に固有の購入パターンが含まれている場合がある。この例では、顧客が権利要求の一部として最近の購入商品リストを提供した場合、組織はそれらに関連する学習データを識別できるかもしれない。
- 学習データで直接または間接的に個人を識別できない場合がある【GDPR条文に比べて不正確？】。これを証明できる場合は、（GDPR第11条により）GDPR第15条から第20条に基づく個人の権利は適用されない。ただし、個人が識別を可能にする追加情報を提供する場合、これはもはや当てはまらず、組織は個人が行う要求に対応する必要がある。
- AIで個人データを利用すると、情報提供、アクセス、訂正、消去、処理の制限、通知に対する個人の権利を履行することが困難になる場合があることをICOは認識している。要求が明らかに根拠のない、または過度である場合、料金を請求するか、要求への対応を拒否できる場合がある。ただし、AIで実行するのが難しいとか、要求の動機が不明確だという理由だけで、明らかに根拠のない、または過剰な要求と見なすべきではない。

第4部： AIシステムにおいて個人の権利をどのように保証するか？（概要）

◆ AIのアウトプットに対する個人の権利要求をどのように保証するべきか？

- 通常AIシステムがディプロイされると、そのアウトプットは個人のプロファイルに保存され、個人に対して何らかのアクションを実行するために利用される。例えばWebサイトでのレコメンデーションは、個人プロファイルに保存されたアウトプットによって駆動されるかもしれない。このデータが個人データに該当する場合、通常すべての権利の対象となる（これらの権利に対する例外やその他の制限が適用される場合を除く）。
- 学習データにおける個々の不正確さの影響は無視できるかもしれないが、モデルの不正確なアウトプットは個人に直接影響を与える可能性がある。したがって、モデルのアウトプット（または入力データ）の訂正要求は、学習データの訂正の要求よりも行われる可能性が高くなる。ただし前述のように、予測は、事実の記述ではなく予測スコアとして意図されている場合、不正確なものとは言えない。個人データが不正確でない場合、訂正の権利は適用されない。
- 個人から提供されたデータのさらなる分析によって得られた個人データは、ポータビリティの権利の対象ではない。すなわち、個人に関する予測や分類などのアウトプットは、ポータビリティの権利の範囲外である。

○個人の権利は、モデル自体に含まれるデータとどのように関連しているか？

- 個人データがAIモデル自体に含まれることもありうる。設計による場合と、偶然による場合の2つである。

◆ 設計により個人データを含むモデルに関する要求にどのように対応するべきか？

- 設計により個人データがモデルに含まれている場合、それは、サポートベクターマシン（SVM）などの特定のタイプのモデルには、ディプロイメント中に新たなサンプルを区別するために、学習データ内のいくつかの重要なサンプルが含まれているためである。これらの場合、個々のサンプルの小さなセットがモデルの内部ロジックのどこかに含まれている。
- 学習データセットには通常、数十万のサンプルが含まれており、モデル内で直接利用されているのはごくわずかである。したがって、関連する個人の1人が要求を行う可能性は非常に低いが、可能性は残っている。
- データ主体の要求がアクセス要求である場合は、モデルを変更せずに対応できるかもしれない。要求が訂正や消去要求である場合、モデルを（訂正後のデータを使用して、または消去されたデータを使用せずに）再トレーニングするか、モデルを完全に削除しないと、対応が不可能な場合がある。

第4部： AIシステムにおいて個人の権利をどのように保証するか？（概要）

◆ 偶然によりモデルに含まれた個人データに関する要求にどのように対応するべきか？

- 一部のモデルは偶然により個人データを「暴露」する可能性がある。このような場合、権限のない者がモデルの動作を分析することで、学習データの要素を復元したり、学習データに誰が含まれていたかを推測したりできる可能性がある。（前述のモデル反転攻撃やメンバーシップ推論攻撃。）
- これらのシナリオでは、アクセス、訂正、および消去の権利を行使して実行することが困難または不可能であるかもしれない。当該個人がモデルから個人データを推測できるという証拠を提示しない限り、個人データを推測できるかどうか、したがってリクエストに何らかの根拠があるかどうかを判断できないかもしれない。
- 偶然的な「暴露」のリスクを最小限に抑えるために、最先端のテクノロジーに照らして、モデルから個人データが推測される可能性を定期的かつ積極的に評価するべきである。

○自動意思決定に関連する個人の権利をどのように保証するか？

◆ 自動意思決定に関連する権利がAIシステムにとって固有の問題になるのはなぜか？

- 機械学習に基づく複雑なシステムは、人間による意味のあるレビューに対して多くの課題を提示する可能性がある。機械学習システムは、データパターンに基づいて人々に関する予測や分類を行う。非常に統計的に正確である場合でも、個々のケースで誤った判断に達することがある。このようなエラーは、人間のレビュー者が識別、理解、修正するのが簡単ではないかもしれない。
- 個人からのすべての異議申立てが意思決定を覆す結果になるわけではないが、多くが覆される可能性があることを予期するべきである。機械学習システムでこれが当てはまる理由は2つある。
  - 当該個人が「外れ値」である場合： 当該個人の状況が、学習データで考慮されていた状況とは大幅に異なる。類似した個人に関する十分なデータでトレーニングされていないため誤った予測や分類を行う可能性がある。
  - AI設計における仮定に異議を唱えることができる場合： たとえば年齢などの連続変数が、モデリングプロセスの一部として、20～39歳のような大雑把な年齢範囲に分割された（「容器に入れられた」）可能性がある。「容器」をより細かくすることで、異なる年齢の人々に対して実質的に異なる予測を行うような、違ったモデルが生成される可能性がある。このデータの前処理やその他の設計上の選択の妥当性は、個人からの異議を受けて初めて問われることとなるかもしれない。

第4部： AIシステムにおいて個人の権利をどのように保証するか？（概要）

◆ 自動意思決定に関連する権利を保証するために、どのような措置を取るべきか？

- 組織は以下を行うべきである。
  - 設計フェーズから人間による意味のあるレビューをサポートするために必要なシステム要件を検討する。特に、人間のレビューと介入をサポートするために設計された解釈可能性の要件と効果的なユーザーインターフェイスを検討する。
  - 人間のレビュー者向けに適切なトレーニングとサポートを設計し提供する。
  - 個人の懸念に対処したりエスカレーションしたり、必要に応じてAIシステムの意思決定を無効にすることができるように、適切な権限、インセンティブ、サポートをスタッフに提供する。
- ただし、注意すべきいくつかの追加の要件と考慮事項がある。ICOとAlan Turing Instituteの「AIを用いた意思決定を説明する」ガイダンスでは、AIシステムが複雑になることで、個人への意味のある説明がどのように難しくなるかをを調べている。またシステムが複雑すぎて説明できない場合は、意味のある異議申立てや、人間の介入、レビュー、別の見解の表明もできなくなるかもしれない。
- 個人が自分の権利を行使するためのプロセスは、シンプルでユーザーフレンドリーでなければならない。たとえば、自動意思決定の結果をWebサイトで伝達する場合、そのページには個人がスタッフメンバーに容易に連絡できるようなリンクや明確な情報を含めるべきである。
- ヒューマンインザループを持つことは、トレードオフを伴う可能性もある。AIが生成したアウトプットを検証や却下するために人間のレビュー者が追加の個人データを必要がする場合、結果的にさらなるプライバシー情報が求められてしまうかもしれない。また、自動プロセスの最後に人間のバイアスが再導入されてしまう恐れもある。

○人間の監督の役割は何か？

- AIを使用して個人に関する法的または同様に重要な意思決定（例えばローン審査や人事採用）を実施する場合、これらの意思決定が適切な人間の監督なしに行われるリスクがある。

- このリスクを軽減するには、人間による監督を提供するようにアサインされた担当者が、システムのアウトプットに関与し、批判的であり、適切な場合はいつでも異議を唱えられるようにするべきである。

第4部： AIシステムにおいて個人の権利をどのように保証するか？（概要）

◆ 完全自動意思決定と部分的な自動意思決定の違いは何か？

- AIシステムは次の2つの方法で使用できる。
  - 完全自動意思決定： AIシステムが自動意思決定を行う。例えばローン申請を自動的に審査するシステム。
  - 意思決定支援： AIシステムは人間の意思決定者を検討過程で支援するのみである。例えばローン申請の許可／却下を決定する際にローン担当者を支援するための追加情報を提供するシステム。
- 完全自動意思決定が、部分的な自動意思決定よりも一般的にリスクが高いか低いかは、個々の状況によって異なる。したがって、個々のコンテキストに基づいてこれを評価する必要がある。
- それらの相対的なメリットに関係なく、自動意思決定は、データ保護法令において人間による意思決定とは異なる方法で扱われる。すなわちGDPR第22条において、個人に法的または同様に重大な影響を与える完全自動意思決定は、より限定的な適法性の根拠に制限され、特別な保護措置を講じることが要求されている。
- 対照的に、意思決定支援ツールとしての使用は、このような制約の対象ではない。ただし人間によるインプットは意味のあるものである必要がある。人間が「めくら印」を押したからといって、意思決定がGDPR第22条の対象外になるわけではない。個人について最終的な意思決定が下される前の人間によるレビューや介入の程度と品質は、AIシステムが完全自動意思決定に使用されているのか、単に意思決定支援ツールとして使用されているのかを決定する重要な要素である。
- 重要な考慮事項は次のとおりである。
  - 人間のレビュー者は、システムによるレコメンデーションの確認に関与しなければならず、システムによるレコメンデーションをルーチン的に個人に適用するべきでない。
  - このようなレビュー者の関与は、単なる形ばかりの関与ではなく、積極的なものでなければならない。それらは、システムによるレコメンデーションに反対する「権限と能力」を含め、意思決定に実際の「意味のある」影響を与えるものであるべきである。
  - レビュー者は、システムによるレコメンデーションを「検討」および「解釈」し、利用可能なすべての入力データを検討し、その他の追加要素も考慮しなければならない。