資料WG12−2

EUにおけるAI規制と偽・誤情報対策

1．AI法案
2．デジタルサービス法
3．いくつかの論点

2024年4月5日

生貝直人　博士（社会情報学）
一橋大学大学院法学研究科教授

1．AI法案：禁止されるAI行為

- 近日中に正式に確定する同法では、AIシステムに対してリスクに応じて①禁止されるAI行為、②ハイリスクAIシステム、③特定のAIシステムの透明性義務、④自主的な行動規範の４類型に分けた規律を行う他、⑤生成AIを含む汎用目的AI（General Purpose AI）に対する特別な規律を設ける
- 禁止されるAI行為（概略）
  - 判断能力を著しく損なうサブリミナル技法や操作的・欺瞞的技法
  - 年齢、障害、特定の社会的・経済的状況に起因する脆弱性の悪用
  - 本人や集団に不利な影響を与える社会的スコアリング（例外有）
  - 個人の性格特性や特徴のプロファイリングのみに基づく犯罪予測
  - 顔識別DB作成目的のインターネット・CCTV顔画像無差別スクレイピング
  - 職場及び教育機関における感情識別（医療・安全目的の例外有）
  - 生体識別データに基づく特別カテゴリーデータの推測（例外有）
  - 法執行目的の公共のアクセス可能な空間での生体識別（例外有）

※本資料でのAI法案の内容は、特に言及が無い限り2024年3月13日欧州議会採択テキストに基づく。
https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html

1．AI法案：ハイリスクAI

- ハイリスクAIのカテゴリー
  - 既存EU法での適合性評価義務対象：機械、玩具、レジャー用船舶、リフト、爆発性雰囲気装置、無線機器、圧力機器、索道設備、個人用保護具、ガス機器、医療機器
  - AI法での新たな指定：バイオメトリクス（遠隔生体識別・感情識別）、インフラ管理・運用、教育や職業訓練での学生や希望者の評価や受入れの合否、雇用、労働管理、自営業へのアクセス、重要な民間・公共サービス（公的支援金給付、融資、緊急対応措置）、法執行、移民・亡命・国境管理、**司法又は民主主義プロセス**
    - ※審議プロセスの中でデジタルサービス法のVLOP/VLOSEが用いるレコメンダーシステムを直接ハイリスクAIに含む提案も行われたが、最終的には含まれず
- ハイリスクAIシステムの要求事項
  - リスクマネジメントシステムの構築、データとデータガバナンス、技術文書、記録保持、透明性と利用者への情報提供、人間による監視、正確性・堅牢性・セキュリティ
  - 要求事項具体化手段としての整合規格
- ハイリスクAIシステム提供者の義務：上記要求事項の遵守確保等
- ハイリスクAI配備者の義務：**基本権影響評価**の実施と当局への提出

1．AI法案：ハイリスクAI

- 付属文書III　8. 司法及び民主的プロセス
  - (b) 選挙や国民投票の結果、または選挙や国民投票における自然人の投票行動に影響を与えるために使用されることを意図したAIシステム。これには、管理上または物流上の観点から政治キャンペーンを組織、最適化または構造化するために使用されるツールなど、自然人が直接さらされないAIシステムの出力は含まれない。

1．AI法案：特定のAIシステムの透明性義務

- 自然人と直接対話するAIシステム提供者の開示義務
- 感情識別・生体識別システム配備者の本人通知義務
- ディープフェイク生成AIシステム配備者：当該コンテンツが人為的に生成・操作されたものであることを開示する義務
- 汎用目的AIを含むコンテンツ生成AI提供者：AIシステムの出力が機械可読形式でマークされ、人為的に生成・操作されたことを検出できることを保証する義務
  - 前文120「本規則において、特定のAIシステムの提供者および配備者に課される、当該システムの出力が人為的に生成または操作されたものであることを検知し、開示することを可能にする義務は、**規則（EU）2022/2065（※デジタルサービス法）の効果的な実施を促進**するために特に関連する。これは、特に、**人工的に生成または操作されたコンテンツの拡散から生じる可能性のあるシステミックリスク、特に偽情報を通じたものを含む民主的プロセス、市民的言説および選挙プロセスに対する実際または予見可能な悪影響のリスク**を特定し、軽減するための、非常に大規模なオンラインプラットフォームまたは非常に大規模なオンライン検索エンジンのプロバイダーの義務に関して適用される。」

1．EU AI法案：汎用目的AI

- 汎用目的AIモデル提供者の義務
  - 設計や学習等の技術文書作成と当局への提供
  - 下流事業者への情報開示
  - DSM著作権指令4条（学習データオプトアウト）遵守措置、学習データ要約
- システミックリスクを有する汎用目的AIモデル（10^25FLOPs以上等）提供者の義務
  - システミックリスク特定・軽減のためのレッドチームテスト実施・文書化を含むモデル評価
  - **システミックリスクの評価・軽減**
  - 重大インシデントへの対応文書化と当局への報告
  - サイバーセキュリティ対策

→それぞれの義務は整合規格により具体化、それまでは欧州委員会主導で策定する行動規範（codes of practice）の遵守

3条(65)「「システミックリスク」とは、汎用目的AIモデルの高インパクト能力に特有のリスクであって、**その影響範囲の広さにより連合市場に重大な影響を及ぼし、または公衆衛生、安全、治安、基本権もしくは社会全体に対する実際の若しくは合理的に予見可能な悪影響**により、バリューチェーン全体にわたって大規模に伝播し得るリスクをいう」

1．EU AI法案：汎用目的AI

- 前文110「汎用目的AIモデルは、重大事故、重要部門の混乱、公衆衛生及び安全に対する重大な影響、 民主的プロセス、公共及び経済の安全に対する実際の、又は合理的に予見可能な悪影響、 **違法、虚偽、又は差別的なコンテンツの流布を含むが、これらに限定されないシステミックリスク**をもたらす可能性がある。（…）」

- 前文133「**さまざまなAIシステムが大量の合成コンテンツを生成できるようになり、人間が生成した本物のコンテンツとの区別がますます難しくなっている。**こうしたシステムが広く利用可能になり、その能力が高まることは、情報エコシステムの完全性と信頼性に重大な影響を及ぼし、**誤情報や大規模なマニピュレーション、詐欺、なりすまし、消費者への欺瞞といった新たなリスクを引き起こす。**こうした影響、技術の進歩の速さ、情報の出所を追跡するための新たな手法や技術の必要性を考慮すると、こうしたシステムの提供者に対し、機械が読み取り可能な形式で表示し、その出力が人間ではなくAIシステムによって生成または操作されたことを検出できる技術的ソリューションを組み込むことを求めることが適切である。(…)」

2．デジタルサービス法とAI

- EUのプロバイダ責任を規定してきた電子商取引（2000年）を元に、違法・有害情報に対するプラットフォームの責任・責務や透明性のあり方を全面的にアップデート
- 媒介サービス事業者（IS）一般やオンラインプラットフォーム（OP）事業者一般に適用される規律の他、EU域内で月間アクティブ利用者4,500万人以上を有する「超大規模オンラインプラットフォーム（VLOP）」＋「超大規模オンライン検索エンジン（VLOSE）」事業者に、偽・誤情報を含むシステミックリスクの評価・軽減義務を課す
  - 2023年4月25日に17のVLOPと2のVLOSEが指定、2024年2月全面適用開始
- デジタルサービス法の要点
  - コンテンツモデレーション：透明性と救済
  - プロファイリング関連規制
  - VLOP/VLOSEとシステミックリスクの評価・軽減

①コンテンツモデレーション：透明性と救済

3条(t)：「コンテンツモデレーション」とは、自動的か否かに関わらず、媒介サービス提供者が行う、特にサービス受領者が提供する違法コンテンツ又はその利用規約に適合しない情報の検出、識別、対処を目的とした活動をいい、降格、収益不能化、アクセス不能化、削除など、違法コンテンツ又はその利用規約違反情報の利用可能性、可視性、アクセス性に影響を与える措置、サービス受領者のアカウントの終了又は停止など、サービス受領者の情報提供能力に影響を与える措置を含む。

- 利用規約へのコンテンツモデレーションポリシー明記（IS、14条）
  - 利用者提供情報に関する制限の情報（**アルゴリズムによる意思決定と人間によるレビューを含むコンテンツモデレーションのあらゆる方針・手順・手段・ツール**、内部苦情処理システム手続に関する情報を含む）
  - 制限の実施における表現の自由やメディアの自由・多元性、その他基本権等の利益への配慮義務
  - VLOP/VLOSEは全サービス提供加盟国の言語で当該情報を提供
- 透明性レポート（IS~VLOP段階、15条他）→VLOPの第一次レポート
  - 当局命令・対応、違法・規約違反別の通知・対応件数と対応時間、コンテンツモデレーション担当者訓練内容、**自動処理のエラー率指標とセーフガード措置等**（※VLOPは加盟国の公用語ごとに整理）
- 削除等の理由の説明（IS、17条）　→欧州委による集約データベース
  - コンテンツ削除・降格やアカウント停止等を受けた利用者への明確かつ具体的な理由説明
- 削除等に異議がある場合の内部苦情処理システム整備（OP、20条）
  - 削除やアカウント停止等の判断が誤っていた場合の回復等
- さらに異議がある場合の裁判外紛争処理の利用（OP、21条）
  - 紛争処理機関に対する当局の認定等

②データ保護：プロファイリング規制

- PF上の**ターゲティング広告**のパラメータ等の明示（OP~VLOP段階、26条他）
- **レコメンダーシステム**のパラメータ明示とユーザーによる修正可能性（VLOPはプロファイリングに基づかない選択肢の提供を含む）（OP～VLOP、27条他）
- **GDPR特別カテゴリー個人データのプロファイリング広告利用禁止（OP、26条3項）**
- **青少年保護と未成年個人データのプロファイリング広告利用禁止（OP、28条2項）**

- ※ダークパターンの禁止（OP、25条）：「サービス受領者を欺いたり操作したりするような方法で、又はその他の方法でサービス受領者が自由かつ情報に基づく決定を行う能力を実質的に歪めたり損なったりする方法で、オンライン・インターフェースを設計、組織、運用しないこと」

③VLOP/VLOSE：偽・誤情報を含むシステミックリスクの評価と軽減

- VLOP/VLOSEは、自らのサービスがもたらしうる違法コンテンツ流布、**基本権**（人間の尊厳、プライバシー、個人データ保護、表現・情報の自由、非差別、児童の権利、消費者保護）、**市民言説と選挙**、ジェンダー暴力・**公衆衛生**・青少年保護等への影響**等の「システミックリスク」を自ら特定・分析・評価し（34条）、合理的・比例的・効果的な軽減措置を採る義務（35条）**と、公共の安全・公衆衛生への重大な脅威における危機対応メカニズムにおいて出される欧州委員会の要請決定の対象となる（36条）
- 欧州委員会が奨励・推進・招請して策定する、**行動規範（codes of conduct）（45条）**や危機プロトコル（48条）**を通じて具体化する共同規制メカニズム**
  - デジタルサービス法採択以前から偽情報行動規範が策定、2022年6月の改訂によりディープフェイク等への対応も含まれる
- 34条・35条の義務及び、行動規範・危機対応プロトコルの遵守について、年1回以上の独立監査を受ける義務（37条）
  - 評価・緩和措置検証のための外部研究者データアクセス提供義務（40条）

デジタルサービス法とAI法

- AI法案前文118「本規則は、AIシステムおよびモデルを規制するものであり、関連する市場関係者に対し、それらを市場に投入し、サービスを開始し、または域内で使用するための一定の要件および義務を課すことにより、規則（EU）2022/2065 （※デジタルサービス法）により規制される、そのようなシステムまたはモデルをサービスに組み込む媒介サービスの提供者に対する義務を補完するものである。そのようなシステムまたはモデルが、指定された超大規模オンラインプラットフォームまたは超大規模オンライン検索エンジンに組み込まれる限りにおいて、それらは規則（EU）2022/2065に規定されたリスク管理の枠組みの対象となる。その結果、規則（EU）2022/2065が対象としていない重大なシステミックリスクが出現し、そのようなモデルで特定されない限り、本規則の対応する義務は履行されていると推定されるべきである。この枠組みの中で、超大規模オンラインプラットフォームおよび超大規模オンライン検索エンジンのプロバイダーは、潜在的な悪用から生じるシステミックリスクだけでなく、サービスで使用されるアルゴリズムシステムの設計がそのようなリスクにどのように寄与するかを含め、サービスの設計、機能、使用から生じる潜在的なシステミックリスクを評価する義務がある。これらの提供者はまた、基本的権利を遵守し、適切な緩和措置を講じる義務がある。」

４．いくつかの論点

- AI法案：AI提供者・配備者等
  - 製品安全
  - プロファイリング
  - 偽・誤情報にとどまらない操作・欺瞞
- デジタルサービス法：プラットフォームレイヤー
  - AIとコンテンツモデレーション、レコメンダーやプロファイリング
  - AI生成コンテンツ流通への対応
- 両法の補完関係
  - AI法案におけるAI生成コンテンツ検出可能化義務と、デジタルサービス法におけるPF側の検知・開示措置（システミックリスク軽減）
  - システミックリスク軽減義務